QUALITY IN PRIMARY CARE ECONOMIC APPROACHES TO ANALYSING QUALITY-RELATED PHYSICIAN BEHAVIOUR

Michael Kuhn

Office of Health Economics
12 Whitehall London SW1A 2DY

ISBN 1 899040 67 6

Acknowledgement

I am grateful to Jennifer Dixon, Bonnie Sibbald and Peter Zweifel for their extremely helpful comments, and to my colleagues at the Centre for Health Economics and Giuliano Masiero for instructive discussions of a variety of the issues involved. I am also indebted to Jon Sussex for his excellent comments and his unrelenting editorial support. All responsibility for remaining errors lies entirely with me. Funding by the OHE is gratefully acknowledged.

About the authors

Michael Kuhn is an economist. He holds a doctorate from the University of Rostock, Germany and is currently a lecturer at the Department of Economics and Related Studies and the Centre for Health Economics of the University of York. His research interests lie in the fields of industrial organisation and incentive theory as applied to health policies and governmental regulation.

THE OFFICE OF HEALTH ECONOMICS

The Office of Health Economics (OHE) was founded in 1962. Its terms of reference are to:

- commission and undertake research on the economics of health and health care;
- collect and analyse health and health care data from the UK and other countries;
- disseminate the results of this work and stimulate discussion of them and their policy implications.

The OHE is supported by an annual grant from the Association of the British Pharmaceutical Industry and by revenue from sales of its publications, consultancy and commissioned research.

Independence

The research and editorial independence of the OHE is ensured by its Policy Board:

Peer Review

All OHE publications have been reviewed by members of its Editorial Board and, where appropriate, other clinical or technical experts independent of the authors. The current membership of the Editorial Board is as follows:

Further information about the OHE is on its website at: www.ohe.org

CONTENTS

FOREWORD

by Professor Hugh Gravelle (National Primary Care Research and Development Centre and the Centre for Health Economics, University of York)

This monograph sets out the insights from applying an economic perspective to a fundamental problem for any health service: ensuring cost-effective care when outcomes from care are uncertain and there is imperfect information about the activities of providers and the factors outside their control which also affect outcomes.

The focus is on primary care but the lessons carry over to other sectors.

The text of the monograph was written before the details of the proposed new NHS contract for British GPs – the General Medical Services (GMS) contract – were known. It is clear that the new contract represents the most substantial change in the way in which GPs are paid since the founding of the UK's National Health Service (NHS) 55 years ago. The fundamental aim of the new contract is to improve the quality of primary care by providing a powerful and wide ranging set of financial incentives.[1] It is therefore instructive to examine the contract in the light of the literature so admirably surveyed in the monograph.

The key new features of the new contract are:

- practice level contract. Primary Care Trusts (PCTs – the NHS organisations responsible for purchasing care for local populations) will contract with practices, rather than individual GPs, for the provision of services;
- levels of contracted service. Practices will have to provide 'essential' services and in normal circumstances will be expected to provide 'additional' services such as child health surveillance and minor surgery. They can also contract to provide 'enhanced' services, such as more advanced minor surgery and flu immunisations or anything else they may agree in negotiation with the local PCT. The payment for each level of service will be related to the number of patients in the practice, adjusted by age, sex and measures of need;
- explicit quality incentives. Practices will receive points according to their achievements on over 140 performance indicators. The funds received per quality point attained will vary with the age, sex and need adjusted number of patients in the practice;
- PCTs will have increased flexibility in commissioning additional and enhanced services.

1 The details of the proposed contract are set out at http://www.nhsconfed.org/gmscontract/

Quality related payments are expected to constitute a larger proportion of practice income than hitherto and, more importantly, practice income will be more responsive to quality. The new contract has much more high powered incentives for quality and there are grounds for thinking that its structure is potentially an improvement on the existing contract.

Since practices are best able to monitor and control the costs of their activities it is appropriate that they bear them and are remunerated by pricing the quality points to reflect the relative value society places on the benefits from these activities. Michael Kuhn, in chapter 6 of this book, describes the hazards of regulating for quality, along with the potential benefits. Direct regulation by setting targets which must be achieved across a range of activities is inflexible and neglects the fact that the costs of achieving the targets may differ across practices. The points pricing system lets practices decide on how to allocate their efforts but still enables the NHS to influence the mix of activities by adjusting relative and absolute points prices and their monetary value. The fact that the contract is practice based and that each practice, however large, need only have one GP, makes it easier for practices to consider alternative mixes of professional skills to increase quality.

The explicit incentives in the current GP contract are directed at a narrow range of practice activities and so run the risk that they divert effort from unrewarded but valuable activities. The considerable increase in the range of quality areas which will be rewarded means there is less danger of inappropriate allocation of effort within practices. The contract also attempts to reward action across a range of

activities directly by providing additional 'holistic' payments for doing well across a wider range rather than specialising in just a few activities to achieve higher payment.

Under the current contract the main incentive for practices to be more responsive to patients is exit: patients may move to other practices thereby reducing capitation-based practice income. This incentive has its limitations, as Michael Kuhn makes clear in chapter 4. The new contract supplements the exit mechanism by strengthening the role of voice[2] by providing financial incentives for practices to survey their patients and to act on the results. Since the surveys cover a range of patient experiences, including patients' views on the ease and convenience of access, and the interpersonal skills of doctors and nurses, practices will be rewarded for a broader set of activities with less tangible but important benefits to patients.

It is a fundamental principle of performance related pay (discussed in section 6.1) that rewards should depend as little as possible on factors outside the control of those being incentivised. The achievement of some of the performance indicators in the new contract depends in part on the patient. For example patients may refuse to attend the practice for review of their condition or may have an allergy to the recommended drug. The new contract will permit exception reporting so that such patients will be allowed for when achievement of performance measures is calculated. Exception reporting will thus also reduce incentives for cream skimming of patients to avoid those who would make it more difficult to earn quality points.

It is possible to argue about many of the features of the new contract, especially about some of the performance indicators and the relative quality points attached to them. But one firm prediction can be made: there will be consequences from the new set of incentives which have not been intended by the negotiators.

2 See Hirschman, Albert O. *Exit, Voice and Loyalty: Responses to Decline in Firms, Organizations, and States.* Cambridge: Harvard University Press, 1970.

There will be light monitoring of the contract by PCTs. Practices will self-report their quality points in an annual report to their PCT which will also visit them. Given the sums involved, there appear to be considerable incentives for gaming and misreporting by practices and, as the 1990 GP contract and the introduction of GP fundholding in 1991 showed,[3] at least some GPs will respond. The challenge will be for those monitoring and revising the contract to ensure that GPs channel their effort in productive rather unproductive directions. It is therefore important that the effect of the contract on GP behaviour be evaluated to assist in the cost-effective reform of quality incentives.

3 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, 375-398.

EXECUTIVE SUMMARY

The importance of primary care physicians in the provision of high quality health care receives increasing recognition by researchers and policy-makers. Physicians working in primary care assemble packages of care for their patients using medicines, inputs from secondary care and inputs from other primary care staff, together with their own diagnostic and therapeutic efforts. They act as agents on behalf both of the patient and of the payer.

This review provides an understanding of quality incentives for primary care physicians and how they are shaped by the organisation of primary care. The main focus is on explaining in an intuitive way the economic mechanisms behind the incentives as well as their policy implications. The arguments are illustrated by examples and by empirical work relating to current policy issues. The focus is on European health care systems with an emphasis, albeit not an exclusive one, on the UK's National Health Service (NHS).

As a basis for economic analysis, a concept of quality is suggested that embraces both the patient's health gain and the convenience attributes of care. Donabedian's distinction between the 'structure', 'process' and 'outcome' aspects of quality provides a useful framework. Thus, physicians produce health outcomes using medical equipment and skills (structure) in a process that combines effort with a range of variable inputs including secondary care and pharmaceuticals. Quality incentives relate to investments in 'structure' as well as to the effort provided and the input mix chosen in the 'process' of care.

Quality incentives are shaped by the institutions of primary care, such as the payment scheme, performance standards and the practice environment. Policy-makers shape the quality incentives by designing institutions. In this, they usually have to trade-off a high quality service against cost-containment and efficiency against equity. The role of physicians as agents on behalf of both the patient and the payer has implications for their incentives and for the policies that shape them.

Primary care physicians can be viewed as maximising a utility function that contains income, professional and social status, intrinsic benefits and altruistic concerns, as well as the cost of effort. Quality incentives can be attached to each of the elements of the utility

function. The sources of these incentives are competition, regulation, the physician's ethical values and professional and social norms.

In as far as physicians maximise income, they trade-off at the margin the revenue generated from the provision of quality against the cost. Quality incentives then increase with the margin between unit fee and unit cost and with the responsiveness of patient demand to quality. The latter increases the more physicians there are around to choose from and the more outside options that patients have apart from the physicians' services. The sensitivity of patient demand to the quality of care diminishes the greater are the costs of switching from one physician to another and the poorer the information they have about the quality of the service.

Asymmetric information about physicians' skills and effort restricts patient choice and stifles competition. Where quality cannot be observed it is likely to be under-provided. Asymmetric information can be resolved by a number of mechanisms:

- by search, if patients can inspect the relevant quality aspect;
- by signalling of hidden information by the physician or the acquisition of a reputation if the service is of an experience nature, i.e. patients are able to determine quality once they make use of the physician's services;
- by collective reputation or professional credentialling by independent experts if the service is of a credence nature, i.e patients cannot determine its quality;
- by regulatory measures to reduce informational asymmetries, including (re-) accreditation, certification and the use of performance indicators.

Generally, the resolution of asymmetric information involves a social cost that should be counted as an indirect cost of providing quality.

The nature of the remuneration and reimbursement system has an important influence on the provision of quality. If the physician is mainly motivated by financial concerns, fixed budgets may entail an under-provision of services and quality. The same applies for a flat salary, which may also induce too many referrals and prescriptions. Capitation gives rise to correct quality incentives for financially

motivated physicians if and only if patient demand is responsive to all of the relevant quality dimensions. Otherwise it might lead to the under-provision of quality dimensions that are unobservable or to discrimination between patients on the basis of quality. Fee-for-service tends to lead to an over-provision of services with ambiguous implications for quality. Quality may be so high as to be cost-ineffective. However, quality may be too low if, for example, physicians were to treat patients themselves even if a referral would be more appropriate. If patients are permitted direct access to specialists, this provides an incentive for physicians to specialise in order to differentiate their services.

While empirical evidence supports some of the theoretical predictions about the incentives provided by different payment systems, little is yet known about the implications specifically for the quality of care.

Recently, there have been moves in some health care systems towards introducing more direct quality incentives into physician remuneration. The advantage of performance pay lies in its provision of direct quality incentives even if patient demand is unresponsive to quality. The design of performance pay schemes is subject to a range of problems relating to:

- the need to equalise reimbursement across all important dimensions of performance;
- provision of team incentives;
- containment of the physician's performance risk;
- determination of the right degree of monitoring;
- determination of performance benchmarks; and
- the regulator's credibility when committing not to extract the physicians' rent by ratcheting up standards.

Concerns about variation in practice sometimes lead to the imposition of best practice guidelines. If physicians differ in their abilities and expertise with particular technologies, an imposition of guidelines may compromise the provision of quality by some physicians. Dissemination of information and encouragement of continuing education may then be preferable to imposition of guidelines.

Altruistic concerns about patients' welfare mitigate the potential for under-provision of quality; but usually not the potential for inefficient resource use. Cost-sharing can induce altruistic physicians to provide optimal levels of quality. Intrinsic motivation, i.e. satisfaction in a job well done, is an important quality incentive for physicians but it can be undermined both by market incentives and regulation. Competition for social status within a peer group or within society in general may provide quality incentives but also exposes physicians to a 'status risk' that should be accounted for in the design of formal performance schemes. Generally, the inter-relationship between non-financial quality incentives and financial incentives provides much scope for future research.

Within many health care systems concerns are voiced that professional self-regulation of quality is inadequate as a safeguard. One possible explanation is that free-riding leads to a lack of incentives to maintain a collective reputation. In the light of this, clinical governance has recently received attention as a potentially powerful mechanism for controlling quality.

Within the UK NHS, Primary Care Trusts are expected to implement national performance standards by introducing a system of clinical governance. Bodies within the Primary Care Trusts supervise the practice of physicians in safeguarding the quality of care. Clinical governance can also be understood as a framework of simple formal and informal rules for (appropriate) behaviour under various contingencies. This facilitates the establishment of reputation by physicians and by the regulator alike. Collective learning and information sharing are understood to be key elements of clinical governance. They can be interpreted as a form of participatory regulation, where physicians are involved in determining their own performance framework. While this makes the regulator's task easier, it also opens a channel for possibly harmful influencing of the regulator.

Primary care physicians play an important role as intermediaries in that they commission secondary care and/or audit its quality. In their role as commissioners they may induce quality competition between providers of secondary care. In their role as auditors they act as

intermediate agents for the regulator with a 'whistle blowing' function. Empirical evidence on the provision of quality in primary care is scarce and in many cases inconclusive. The problem of finding good measures for quality seriously impairs all empirical work, but if anything this demonstrates the remaining scope for empirical research.

1 INTRODUCTION

For some time now, the quality of health care provision has been high on the agenda of health care professionals, policy makers, the public, and researchers alike. Whereas a strong concern with quality is something very natural for a service as fundamental as health care, quality is not a straightforward matter. It can be seen both in an individual context, i.e. as the quality of care received by a patient, and in a societal context, i.e. as the quality of care experienced by a population.

The quality of health care provision depends on the resources available within a health care system. Resource constraints give rise to the two archetypal economic issues of efficiency and equity. Efficiency requires that the best possible quality outcome is generated from a given set of resources, while equity requires that quality of care does not vary too much across patients. Efficiency and equity of health care provision are determined by the design of a health care system and the behavioural incentives faced by those actors who decide on the use of resources in the administration of care.

Recently, policy-makers and researchers have increasingly focused on the role of primary care in the process of resource allocation. Administrators of health care systems as varied as the UK's National Health Service (NHS), the US market based systems and the German social insurance system have recognised the pivotal position of primary care in the assurance of efficiency and quality in the delivery of care (e.g. Oxley and MacFarlan 1994; Saltman and Figueras 1997, chapter 6).

General medical practitioners (GPs) and other primary care professionals take on two important functions. In commonly being a first point of contact for a patient with the health care system, they bear particular responsibility as the patients' agents in ensuring that they receive appropriate care. Secondly, by making decisions on treatment, referrals and prescriptions, they act as 'manufacturers', who assemble care from different primary care, secondary care and pharmaceutical inputs. In this role of 'intermediary', GPs bear responsibility in controlling the quality not only of the primary care inputs they and the other members of the primary care team provide, but also of the secondary care and pharmaceutical inputs they bring in.

Furthermore, in administering care, GPs also act as society's agents, bearing a responsibility to ensure an efficient and equitable use of health care resources. The agency role of primary care physicians and their role as assemblers of care render them pivotal actors in the health care system. A debate on the role of primary care has been rekindled in recent years (e.g. Pringle 1998; Bloor et al. 2000).

This book sets out an extended overview of the economics of quality in primary health care. A reflection on the ongoing policy debate brings forth a multitude of wide-ranging issues, including:

- what incentives drive the provision of quality by primary care physicians?
- how does the institutional structure of primary care affect the quality of care provided to individual patients and across patients?
- what is the effect of physician remuneration on quality?
- what are the implications of practice organisation for the provision of quality?
- in what way can primary care physicians bring an influence to bear on the quality of secondary care, and how does this depend on the institutions in place?
- who is shaping the institutions, such as the reimbursement system or the arrangements for regulation of quality, and to what effect?

These questions are asked from a positive perspective, and the answers will help us to understand the way in which the properties of existing health care systems shape the incentives for the provision of quality.

In order to measure the scope for improvement in the performance of health care provision, we have to establish what constitutes the optimal provision of quality and what factors determine it. To guide policy-making we can then assess whether, and if so how, under- or over- provision of quality may be mitigated by regulatory measures or changes in the institutional design, or more specifically:

- can outcome-related reward provide incentives for the enhancement of quality?
- can the publication of performance indicators enhance patient choice and, thereby, improve physicians' quality incentives?
- can an improvement of quality in the overall system be achieved by integrating or co-ordinating the provision of primary and secondary health care?

1.1 Primary care in European health care systems: some institutional background

To establish a reference point for much of the subsequent argument, I shall outline some of the features of UK and other European health care systems, which will turn out to be important determinants of GPs' behaviour.

According to their funding mode, European health care systems have traditionally been divided into those based on social insurance (Bismarck model: Germany, Austria, Belgium, France, Luxembourg and the Netherlands) and those that rely on taxation (Beveridge model: UK, Denmark, Greece, Ireland, Italy, Norway, Spain, and Sweden).[4] All of these systems have in common that patients are generally exposed to no, or at worst modest, (co-)payments for care services. Of greater concern for our purposes will be the mode of physician remuneration and the degree of patient choice, as emphasised by the scope to switch GPs or the degree of direct access to specialist treatment. Table 1.1 provides a rough classification of some European health care systems according to these two criteria. The table reflects the state of the systems in the early to mid-1990s and, therefore, does not claim to be up-to-date. The purpose of the table is to illustrate the variety of approaches adopted.

4 The one exception to this pattern in Western Europe is Switzerland, which is modelled according to the US system and based on private insurance. Most Eastern European countries are in transition towards social insurance-based systems (Saltman and Figueras 1997, chapter 4).

Table 1.1 Physician payment system and patient choice in European primary care

Country	Mode of reimbursement	Gatekeeping	Patient cost sharing
Austria	Fee-for-service	No	20% of population pays up to 20%
Belgium	Fee-for-service	No	Self-employed pay full cost
Denmark	28% capitation; 63% fee-for-service; 9% practice allowance	Yes (direct access if patients accept significant co-payments)	None (unless direct access)
Finland	Salary	Yes	At low level
France	Fee-for-service; salary in health centres	No	25%
Germany	Fee-for-service	No	None
Italy	Capitation (age-differentiated); some fee-for-service	Yes	None
Netherlands	Fee-for-service for high income patients; capitation (age differentiated) for low income	Yes	None for low income
Spain	Salary; capitation	Yes	None
Switzerland	Fee-for-service; some insurers also pay capitation	Yes	Depending on insurance contract
United Kingdom	Capitation (age-differentiated); some fee-for-service; practice allowance and target payments	Yes	None

Source: Adapted by the author from Rochaix (1998), Table 8.2.

Capitation, i.e. payment according to the number of the patients on a GP's list, practice allowances and salary are all prospective payments in the sense that they are fixed before the physician determines the level of services provided. In contrast, fee-for-service (FFS) and target payments are linked retrospectively to the level of service provided. Gate-keeping refers to an arrangement under which patients do not have direct access to specialist services and have to rely on referrals. Despite variations in the design of payment systems, social insurance systems tend to rely on FFS remuneration and to allow patients direct access to specialists. In contrast, national health services rather rely on prospective payments and a gate-keeping function for GPs. In systems with direct access to specialists, there is usually a smaller share of physicians working in general practice.

The UK was unique in its substantial use of 'fundholding', which from 1991 to 1999 was instituted on a voluntary basis at practice level. Under this fundholding system, a practice received a budget, which it would use for the purchase of secondary care and pharmaceuticals for its patients. The idea was to pass purchasing responsibility on to primary care physicians in order to render them cost conscious in the delivery of care. This system was abolished in 1999 and replaced by a form of fundholding at multi-practice level in the different guise of the budget and commissioning responsibility of the newly formed Primary Care Trusts, each of which typically contains around 30 GP practices.

It is easy to identify some fundamental problems associated with each of the pure forms of payment systems. Prospective payments tend to lead to the under-provision of costly services and of effort by the GPs themselves, and thus to potentially low levels of quality. In contrast, FFS may induce the over-provision of services, and gives insufficient incentives for efficient resource use. Salaries provide no direct incentives either for efficient resource use or for the delivery of high quality care.

Both prospective and retrospective payment are likely to give rise to distortions in the service structure, with implications for quality. Capitation may lead to excessive referrals as a means of shifting costs to secondary care providers. Fundholding may lead to insufficient

referrals if this allows the practitioner to save funds. FFS may lead to fewer referrals as physicians try to increase their income from fees.

The incentives and disincentives arising from the payment systems are reinforced by those arising from the arrangements regarding access to specialist care. Under gate-keeping, which is frequently coupled with registration on a GP's list, there is limited scope for patients to seek alternative providers when dissatisfied. Thus, there tends to be too little patient driven competition between GPs. In contrast, direct patient access to specialists may induce GPs to engage in over-provision of services demanded by fully-insured patients, to encourage them not to go directly to a specialist. There is also an incentive for practitioners to specialise, leading to a potentially inefficient supply of specialist care.

The archetype health care systems, thus, face almost converse problems: national health services tend to struggle with quality problems and insufficient supply of services (micro-inefficiency), while insurance based systems battle against an explosion in expenditure (macro-inefficiency). Quite naturally, recent health care reforms have, therefore, witnessed a move to mixed payment systems, which try to combine elements of both prospective and retrospective payments (Rochaix 1998). The necessity of containing expenditure within insurance systems is also reflected in the presence of cost sharing by patients. Further reforms address patients' direct access to specialist care.

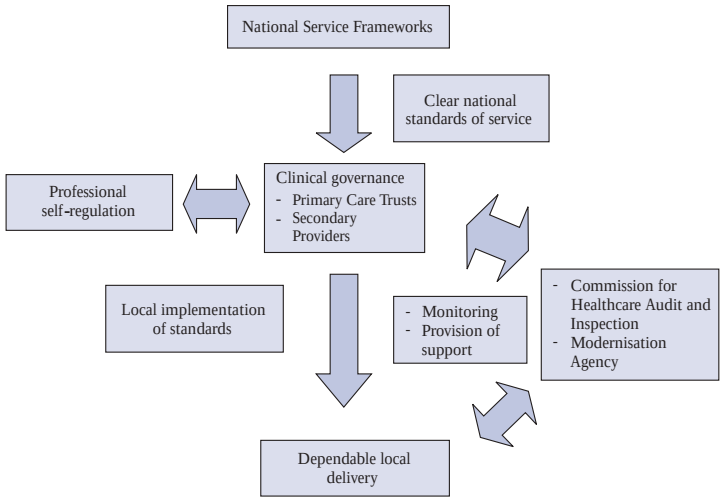
Reforms in Germany have sought to enhance the role of general practice, in general, and encourage the voluntary introduction of gate-keeping (European Observatory on Health Care Systems 2000). Recent proposals go beyond this and suggest a compulsory use of gate-keeping. In contrast, policy initiatives in the UK focus on encouraging a greater degree of specialisation of primary care doctors (Department of Health 2000a, 2003). Reforms of the payment and gate-keeping system affect quality of care and will be topic of this review (chapters 4 and 5).

However, a range of recent reforms and proposals for reform relate more directly to the provision of good quality care. The most comprehensive and far-reaching steps towards guaranteeing the

provision of quality have been taken in the UK. Within the new NHS performance framework (Department of Health 1998, 1999, 2000a), performance standards relating to the provision of high quality care are being set at a national level ('National Service Frameworks') but their implementation is left to lower-tier primary and secondary care organisations.

GP practices have been grouped, along with other community based health services, within Primary Care Trusts. These are to implement a system of clinical governance to improve the quality of care across their constituent practices with a view to meeting the national quality targets. They are also expected to play a role in monitoring and enforcing the quality of secondary care providers. Ultimately, the Primary Care Trusts are accountable to the Department of Health regarding the achievement of the targets specified in the National Service Frameworks. Attainment of the targets, or progress towards them, will be monitored by the Commission for Healthcare Audit and Inspection and additional support is provided by the NHS Modernisation Agency. The structure of this framework is summarised in Figure 1.1.

Figure 1.1 The quality framework within the UK NHS



The reforms embrace a substantial number of the issues this overview reflects upon, including:

- the role of accreditation and continued professional education (section 4.4 of this overview;
- the role of performance standards and (financial) mechanisms to attain them (chapter 6); and
- the roles of self-regulation and clinical governance in quality assurance (chapter 8).

Furthermore, there is an issue about how primary care should be organised and how it should relate to secondary care in order to guarantee the provision of quality (chapter 9).

The recent report by the Advisory Council for Concerted Action in Health Care (Sachverst‰ndigenrat) contains a number of proposals relating to encouraging the provision of quality in German ambulatory care and general practice (Sachverst‰ndigenrat 2001). They embrace similar issues, such as the use of performance indicators; a strengthening of the competencies of the Association for the Promotion of Quality Assurance in Medicine; the provision of better information on quality to patients (addressed in sections 4.3 and 4.4 of this book) in the form of report cards; and the use of quality related pay. Furthermore, it is proposed that the formation of quality circles is encouraged within general practice and that research is focused on the development of clinical guidelines for primary care.

1.2 Scope and outline of the review

The aim of this book is to present an overview of salient issues in the provision of quality in primary care by combining various strands of economic literature. Many of the relevant sources come from health economics. However, it is extremely helpful in understanding the nature and scope of the issues to draw together insights from a variety of other disciplines, including: industrial organisation (dealing with imperfect competition in various types of markets as well as the organisation of firms); regulatory economics; theory of incentives and mechanism design; and managerial economics; as well as an integration of sociology and economics.[5] In compiling this review I

have tried to trade-off the integration of insights from outside health economics against completeness of the literature reviewed. The choice of the literature and of the issues addressed is, therefore, to some extent eclectic. Although the focus is on economics, some references are included from the health services and medical literature in order to provide a background.

This review deals with quality related physician behaviour and how it is shaped by the organisation of health care. The context is the primary care sector within European health care systems. This implies a number of exclusions, which were necessary to narrow down the scale and scope of this review. Firstly, the issue lies with primary care physicians rather than with hospital physicians.[6] Secondly, within the domain of primary care, attention is restricted to physicians and so ignores other important groups of primary care actors such as nurses. Thirdly, the focus on the European context leads to the exclusion of some issues that are salient to the US, such as the role of price competition between physicians and the institution of managed care.[7]

This said, a number of qualifications are in place regarding the inclusion of certain issues. Firstly, from an economic point of view, the quality incentives faced by GPs as providers of primary care are in many cases very similar to those faced by secondary care providers. Thus, despite differences in institutional detail, many of the arguments and insights reviewed generalise to the broader remit of quality in health care.

Secondly, even from an institutional point of view the argument is not narrowly confined to the area of primary care. In those health care

5 For a comprehensive introduction to health economics with extensive sections on physician behaviour and incentives see Zweifel and Breyer (1997), McGuire (2000) and Scott (2000). For an excellent introduction to industrial organisation see Cabral (2000), and for an application to health care markets Dranove and Satterthwaite (2000). For an introduction to managerial economics and the theory of incentives see Milgrom and Roberts (1992). Chalkley and Malcomson (2000) review applications of regulation theory to health care markets.

6 For reviews of the economics of secondary care see Dranove and White (1994), Chalkley and Malcomson (2000), and Dranove and Satterthwaite (2000).

7 For recent reviews of the economics of physician behaviour, which are written predominantly from a US perspective, see Gaynor (1994) and Dranove and Satterthwaite (2000). Glied (2000) reviews the literature on managed care.

systems that allow direct patient access to ambulatory specialist services, such as Germany or France, there is less of a clear separation between primary care as delivered in general practice and secondary care as delivered in hospitals. Here a distinction may rather be drawn according to whether care is delivered in an ambulatory or a hospital context. A number of important arguments that fit naturally into the analysis relate to specialisation by physicians working outside hospitals, which I have included.

Thirdly, as we will find, a concept of quality in primary care that focuses on the treatment administered by primary care physicians themselves is too narrow. It also needs to embrace the other elements of care that are assembled by the GP, involving decisions relating to referral and prescription. The inclusion of these topics implies an extension of this review to issues such as the economics of referrals and prescribing as well as to the interrelationship between primary and secondary care.

The main objective of this review is to provide an intuitive understanding, from an economic viewpoint, of the incentives and institutions that drive the provision of quality in primary care. The second objective is to draw out some general policy implications and illustrate them by referring to some empirical evidence. It is not the aim of this review to provide specific policy advice for any one health care system, which would require a much more focused and detailed level of analysis.

Economic modelling requires a certain amount of abstraction from institutional detail in order to gain an analytic understanding of the incentives and relationships between economic agents. One might criticise the insights generated from economic models for being partial and disregarding the context. For instance, one might criticise a model of income maximising physicians for disregarding the ethical and social motivation that also bears on their behaviour. This should not keep the analyst from distilling insights on financial incentives given the ethical and social motivation. It is then the role of empirical analysis to test whether financial incentives are relevant and whether they accord with theoretical predictions. In this way, the analyst can generate a range of insights about relevant economic incentives.

It is a shared belief amongst economists that while the 'partial' nature of their insights is less than ideal, it is still preferable to attempting a comprehensive 'model' of reality that will turn out to be descriptive and void of analytic content beyond the trivial insight that 'the world is complex'. The understanding is of course that economic insights should be presented in the light of any restrictive assumptions made and that they need to be put into a wider context once they are used in policy making. The present review shares this understanding but for the sake of presentational ease it will usually remain tacit.

The review is organised as follows. The next chapter develops a concept of quality in primary care. Chapter 3 addresses income, status, and intrinsic psychological benefits as determinants of physician behaviour, and discusses how competition, regulation, values and norms shape quality incentives.

Chapter 4 deals in detail with income related quality incentives. In particular, it addresses the conditions under which competition can provide incentives for the provision of quality. In this context, the problem of asymmetric information and private or regulatory means to resolve it are discussed, as well as the problem of discriminatory quality provision. Building on this, chapter 5 then discusses the role of the payment system and how it bears on quality incentives.

Chapter 6 turns towards more direct regulatory measures. Specifically, it addresses the conditions under which performance pay can be used to regulate quality. The second part of this chapter is devoted to the difficulties in using direct performance targets in an attempt to reduce presumed variations in the quality of care.

Chapter 7 deals with non-financial incentives, which, according to common opinion, play a greater role for physicians than for many other professions. In particular, my review of the small but important literature on intrinsic motivation and status competition shows that non-monetary incentives can be important factors behind the provision of quality. Furthermore, they are likely to interact with regulatory or market incentives in an offsetting way, thereby placing a further caveat on regulatory intervention. In the light of these difficulties, an alternative way of securing quality in primary care may lie in allowing the profession to regulate itself. While in the past this

has been the mode of governance within most health care systems, a more controlled approach is now pursued in the form of clinical governance. Chapter 8 briefly explores some economic underpinnings for the role of clinical governance.

Chapter 9 turns to the impact on provision of quality of the organisation of primary care. The first part of the chapter addresses the horizontal aspects of scale (practice size and number of GP partners) and scope (the range of different activities). Here, issues of risk sharing between partners and the strength of formal quality incentives, social interaction and learning advantages from specialisation all play a role. In the second part of the chapter, the role of primary care in the total health care production process is addressed. Here, the conditions under which GPs assemble care and under which they act as effective 'auditors' of the quality of secondary care are important. We also address the issue of co-ordination between primary and secondary care providers and its effect on quality. Chapter 10 concludes.

Summary

- The importance of primary care has recently received increasing recognition by researchers and policy-makers. This is because primary care physicians have an important influence on the quality of health care in assembling a package of care using inputs from secondary care and medicines together with their own diagnostic and therapeutic efforts. In this they act as agents on behalf both of their patients and of the payer.
- It is important to understand the quality incentives facing primary care physicians, as well as how they are shaped by the organisation and regulation of primary care.
- European health care systems can be divided crudely into two groups according to the mode of remuneration of primary care physicians and the presence or absence of a gate-keeping function. Social insurance systems tend to involve retrospective reimbursement (FFS) and no gatekeeping; whereas national health service systems tend to involve prospective payment (capitation, practice budgets) and gate-keeping.
- FFS may stimulate over-provision of services and possibly of quality, from a social point of view, whereas prospective payment may err towards insufficient quality provision.
- Recent reforms or proposals for reform in the UK and Germany aim both at improving incentives for the provision of quality and at providing frameworks for quality control.
- The quality framework for the UK NHS involves the setting of performance standards at national level and their implementation at local level by way of a framework of clinical governance.

2 CONCEPTUALISING QUALITY IN PRIMARY CARE

The scientific literature concerned with the quality of (primary) health care has produced a multitude of concepts and definitions. They range as widely as the underlying disciplines and paradigms, the aims of the studies, the actors involved, and the particular contexts. Greenhalgh and Eversley (1999) caution us to take a holistic view rather than develop a unifying paradigm, which will necessarily be flawed in some, or even most, contexts.

Such a view has led to a number of very general definitions. As one representative example, consider the definition of quality proposed by the Institute of Medicine in 1990, according to which quality represents the "degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge".[8] Besides an acknowledgement of the inherent uncertainty in the successful provision of quality, this definition embraces potentially contradictory aspects, namely:

- preferences ('desired'), applied both at individual and societal level; and
- technical quality, e.g. best-practice based on outcomes research.

According to this definition, quality revolves around a trinity of technical quality, the individual patient's welfare and societal welfare and embraces a number of trade-offs. A patient's perception of quality may deviate from best-practice if that involves a treatment the patient dislikes for its side effects, for its riskiness, or for the time over which health benefits accrue.[9] Societal preferences have to take into account the benefits from the alternative uses of limited resources. Thus, the requirement of cost-effectiveness may at the same time rule out the technically best treatment as well as a treatment preferred by the patient. Blumenthal (1996) provides an instructive and more detailed discussion of these trade-offs as faced by the physician as medical decision-maker. From an economic point of view, this reflects the role

8 As cited in Blumenthal (1996). The same source contains a number of other definitions.

9 For a discussion of these issues and their implications for medical cost-effectiveness studies see Garber et al. (1996), who also provide further references.

of the physician as agent of two principals: the payer – ultimately society – and the individual patient.

In the following, I devise a concept of quality in primary care, which can serve as a backbone for the remainder of this work. Consider primary care provided by a GP to a representative patient over a year, say, and suppose we can express the quality of the provision by some index. This could be a measure of the health gain of a patient undergoing this treatment, for example, the increase in quality adjusted life years (QALYs) relative to the (hypothetical) health state the same patient would experience in the absence of primary care.[10] Alternatively, it could be a more comprehensive measure of quality including both the change in health status and the convenience attributes of care. Indeed, such a general measure of quality is recognised by the medical profession, which includes both the technical aspect of care, as ultimately reflected in the change of the patient's health status, and the quality of the physician-patient interaction (Blumenthal 1996).

If the quality index involves a measure of health status or health improvement one has to be careful to construct it in a way allowing a separation of the determinants relating to the GP from those relating to the patient, as well as from random influences. While the patient's type (e.g. susceptibility to treatment), health status prior to receiving services and health related behaviour all bear on the health gain, these influences lie to great extent beyond the control of the GP. Clearly any measure of the physician's impact on health, i.e. the contribution from the service provided, has to be adjusted to take into account patient type and behaviour. Random influences have to be accounted for by considering expected rather than observed values of the quality index.

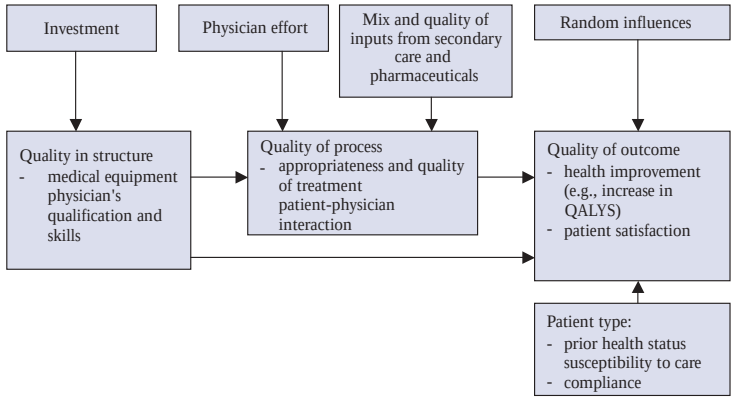
2.1 Quality in the production of primary care

Let us consider those determinants of quality, which relate to the physician. GPs produce care from a bundle of services they provide

10 For an introduction to the concept of QALYs, see Gold et al. (1996) and Zweifel and Breyer (1997, section 2.4).

themselves, a bundle of services they obtain from secondary care providers, and a bundle of pharmaceuticals they prescribe. In this regard, the GPs can be viewed as manufacturers, who assemble a product by combining a range of inputs they procure from upstream providers with a range of inputs they provide themselves.

Figure 2.1 Schematic representation of quality in the production of primary care



In identifying the determinants of quality it is convenient to use Donabedian's distinction between structure, process and outcome (e.g. in Brook et al. 1996). This is illustrated in Figure 2.1. Here, 'outcome', e.g. an increase in QALYs, is produced under a capital 'structure', as given by the physician's stocks of skill and medical capital, and by means of 'process', as given by the variable input choices and the GP's own effort. We have noted already that outcome is not only a function of 'structure' and 'process', which in principle are under the physician's control, but also of the patient's type as well as random influences. Finally, as I will argue shortly, the number of

cases treated by the physician is likely to exert some influence on the quality both of the structure and of the process.

The quality of care increases with the GP's medical skills as well as with the extent and quality of medical equipment (the structural aspect). The GP's medical skills play a two-fold role. First, they enhance the quality of the inputs that the GP provides in person. For any level of time input, a more skilled GP achieves a greater health increase. Second, a skilled GP is in a better position to determine the optimal treatment for a patient and identify the corresponding combination of inputs. The level of medical technology enhances the productivity of the GP's own inputs in producing care. For example, diagnostic equipment improves the GP's decision making and hence the composition of care.

From an economic perspective, both medical equipment and skills constitute capital stocks that are fixed in the short term but can be accumulated over time by making appropriate investments. The extent to which the physician can influence the stock of equipment depends on the nature of the health care system, which may or may not allow the GP to make the necessary investments. Physicians exert a more immediate influence on investments in their own skills. Nonetheless, at least initially, skills are largely determined by the curriculum of medical education. During later stages of their careers, GPs have the scope to update their skills both through learning by doing and participation in programmes of continuing professional development.

Greater effort, like greater skill, improves the effectiveness of a physician's personal interventions and improves the accuracy of diagnosis and treatment decisions. The main difference between effort and ability lies in the fact that the medical skills can only partially be influenced by the GP – and then only by way of long-term investments – whereas effort can be chosen freely and adjusted in the short-term.

The effect of the input choices on the quality of care is determined by two distinct factors: the quality of the respective inputs and the mix of inputs. While quality increases with the quality of each input, the impact of the input mix is less straightforward. First, depending on the patient's characteristics and the nature of the inputs, increasing an

input beyond some boundary level may reduce a patient's health. Examples include over-prescription of pharmaceuticals or unnecessary hospital referrals, which raise the patient's risk of acquiring infections and may reduce their quality of life. Second and more important, while a patient's health may generally be improved by a variety of different combinations, resource constraints make it desirable to find the efficient mix. This is characterised either by the combination of inputs which for a given volume of resources yields the greatest health gain, or alternatively, by the combination of inputs which minimises resource use in achieving a given health gain.

Patients differ with regard to their overall health status, their susceptibility to the level and form of treatment and their preferences about the treatment they are to receive and its mode of delivery. Differences in susceptibility to health care do not only imply that the health outcome varies with a patient's type. They also mean that the marginal effects on health outcome of the GP's effort and other primary care, secondary care and pharmaceutical inputs will differ between patients. Hence the optimal level and mix of inputs will vary between patients.

Differences in patients' initial health status will also lead to different values being accorded to health care quality. For example, a relatively healthy patient is likely to value any given health improvement less than a sicker patient would. Furthermore, patients are likely to vary in their assessment of different modes of treatment. For example, a patient may have a strong aversion against risky surgery. This patient may then rank a treatment involving surgery as a secondary care input lower than an equivalent treatment being based on pharmaceutical inputs. In contrast, a risk neutral patient may prefer the surgery if this avoids negative side effects of pharmaceutical consumption.

These examples illustrate the subjective element of quality, and it is easy to see that a conflict may arise between technically optimal care and what a patient perceives as good quality care. As I will show later, patient heterogeneity with regard to true or perceived quality has strong implications both for GPs' incentives and for the socially optimal administration of care.

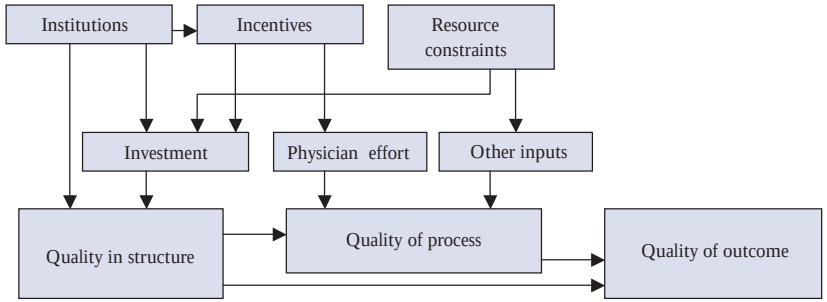
2.2 Incentives, institutions and resource constraints

So far, we have focused on the role of quality in the production of primary care by an individual physician. From this, we have identified aspects of quality relating to structure and to process that translate into the quality of outcomes. We have seen that the physician can influence the quality of structure by making appropriate investments and on process by exerting effort. We have, however, not yet addressed the crucial question as to the factors that determine physicians' choices, and the incentives and the resource constraints the physician faces (see Figure 2.2).

Generally, physicians face both non-monetary and monetary incentives, the former arising from personal ethos or professional norms of conduct as well as from regulatory control, and the latter arising from the payment system. In optimising their objectives physicians are subject to resource constraints, which will always place a bound on the provision of quality. The level of quality that is attainable given these constraints depends both on incentives and on the institutional context. A poorly motivated physician will produce lower quality care with the same set of resources than a highly motivated one.

The role of 'institutions' is also important in determining the quality of care. Under the heading of 'institutions' I include the GP's organisational context (e.g. whether the GP works as part of a practice team or single-handedly), and the formal and informal rules which are applied (e.g. the payment system and the presence of standards of care, but also the presence of social norms). Institutions bear on the physicians' incentives. This is most obvious for the type of reimbursement (discussed in chapters 4 and 5 below) and more direct forms of quality regulation (addressed in chapter 6). But it also relates to social incentives (section 7.3) and group based institutions such as systems of clinical governance (chapter 8). Institutions also influence how primary care is organised, which has its own implications for quality (chapter 9).

Figure 2.2 Role of institutions, incentives and resource constraints in the production of quality care



2.3 Empirical evidence on quality determinants

Medical and health service research has produced a significant literature on the determinants of the quality of care.[11] Rather than review all of this extensive evidence, I shall focus on one recent study (Campbell et al. 2001b) in order to illustrate how the impact of elements of structure and process on outcomes can be tested.

In contrast to most of the previous work that focuses on a single dimension of quality, Campbell et al. (2001b) consider a range of indicators relating to the different aspects of quality in English primary care practices.[12] Specifically, they consider:

- the clinical quality of chronic disease management (angina, asthma in adults, and type 2 diabetes);
- the quality of preventive care (rates of uptake for immunisation and cervical smear);
- access to care;
- continuity of care; and
- interpersonal care;

as dependent variables. On these were regressed the following independent variables:

- practice size;
- routine booking intervals for consultation (5, 7.5 or 10 minutes);
- overall team climate; and
- the deprivation score for the local population.

The results indicate that consultation time was significant in explaining variation in the indicators of clinical care. Practices with 10 minute booking intervals achieved scores for care of asthma, diabetes and angina that were respectively 67%, 21% and 17% higher than those achieved by practices using five minute booking intervals.[13] The

11 For the summary of a recent systematic review see Seddon et al. (2001).

12 They consider a stratified random sample of 60 general practices in six areas of England.

13 This relationship identifies the physician's time/effort input per case as an important process factor in the production of clinical quality. However, nothing is stated about the total time input of physicians in the different practices. If total time input of physicians is the same across practices but practices differ in case load, then the positive relationship between consultation duration and clinical quality indicates a negative relationship between case load and quality.

effects of practice size were ambiguous. Larger practices scored better in diabetes care, whereas smaller practices were superior in access scores. Team climate was significant in explaining higher scores for diabetes care, access, the continuity of care, and interpersonal care.[14] Finally, variations in preventive care were to a significant extent explained by differences in deprivation scores.

2.4 Policy concerns: efficiency and equity

From a positive perspective, the economist seeks to explain physicians' choices of quality as a response to the incentives provided by real-world institutions. From a normative position, the economist asks what combination of inputs, investment and effort, optimises quality for a given set of resources. In a second step and taking into account real-world constraints relating, for instance, to information, contract-writing or policy-making, the economist asks how institutions should be designed to implement a solution which is as close to the optimum as possible. This establishes a benchmark against which to measure the actual institutions and the quality incentives they provide.

When determining what constitutes an optimum, the economist – and later the policy-maker – has to focus on issues of efficiency and equity. The concept of efficiency can be applied at various levels:

- productive efficiency requires that a physician provides any given service at the lowest possible cost;
- allocative efficiency can be understood at two levels:

(a) the physician uses a given set of resources to produce a range of services for a single patient so as to maximise a patient's (health) benefit. This implies an optimal mix of services provided to a single patient and requires that the ratio of the marginal benefits to the patients of different services equals the ratio of their marginal costs;

14 The authors caution that the causal relationship between a good team climate and high outcome scores is unclear. It is plausible that working in a successful practice is less stressful for individuals and thereby fosters a better team climate.

(b) the physician uses a given set of resources to produce a set of services for a patient population so as to maximise the health benefit for the population. This implies an optimal spread of services across patients and requires that the ratio of the marginal benefits of services provided to different patients equals the ratio of the marginal costs of providing these services to these patients. Here, the aforementioned differences between patients become important.

So far, we have considered allocative efficiency at a micro-level, i.e. at the level of an individual decision-maker such as a GP or a primary care practice. In so doing, we have ignored that the provision of primary care services is embedded within a health care system, and even more generally within an economy. While we have derived the optimal level of quality for a given level of resources devoted to primary care, we have not yet addressed the question, as to what level of resources should be devoted to primary care. This raises the issue of allocative efficiency at macro-level:

- macro-efficiency requires that the primary care decision-maker receives a set of resources such that the (marginal) benefits at the micro-level balance the benefits of alternative uses of these resources, e.g. in secondary-care, in other areas of public spending or, indeed, in private investment or consumption. In this regard, it should be noted that while macro-inefficiency may lead to (undue) increases of quality in the short run, the waste of public resources might compromise quality in the long-run.

Thus, the regulator is not only concerned about the incentives for GPs in producing quality, but also about the issue of how much quality should be produced given the limits on public expenditure. As we will see, in answering this second question, the regulator will have to take into account not only the direct benefits and costs of quality provision but also indirect costs of quality, as they arise under asymmetric information and from regulation itself.

The simultaneous requirement of micro- and macro-efficiency confronts the policy maker with an important trade-off between guaranteeing the provision of a high quality service to a patient or a group of patients and economising on public funds. This has

important implications for the role of physicians in the provision of health care and for the incentives to which they should be subjected.[15]

Physicians act as agents on behalf of their patients and this role is particularly pronounced for GPs, who should ideally care for all aspects of their patients' health over an extended period of time.[16] Here, the issue is whether GPs face the proper incentive to provide care in a micro-efficient way. However, micro-efficient behaviour vis-‡-vis their own patients does not imply macro-efficiency. In fact, a dominant concern for their own patients' well-being may induce physicians to expend resources on small increases in their patients' health which are by far outweighed by the foregone benefits from alternative uses of the same resource. In this respect, the GPs act as agents on behalf of the payer for health care, which may be a purchasing agency, the government, or society in general. The role of physicians as agents of two principals (the patient and the payer) with diverging interests (high quality as opposed to cost-effective quality) creates an obvious problem for the structure of incentives they should be given (Blomqvist 1991).

Most policy-makers are concerned not only with efficient provision of health care but also with equitable provision. Since patients differ in their susceptibility to treatment or in the cost of administering care to them, an efficiency-equity trade-off is likely to arise. Suppose, for example, that two patients receive the same marginal benefit from some treatment but that they differ in the marginal cost of treating them. In this case it is efficient to administer less care to the patient who is costlier to treat; but it is obviously not equitable. As agent of both patients and payer, and in their role of assembling packages of care, GPs are likely to be confronted with this trade-off frequently at a practice level of decision-making. The efficiency-equity trade-off arises also in the process of resource allocation to different geographical providers and, in a less direct way, in the allocation of resources between primary and secondary care.[17]

15 Note that the allocative efficiency at micro-level type (b) includes an element of macro-efficiency in that it requires the GP to trade-off the levels of quality supplied to different patients.

16 See Arrow (1963) for an early statement of the physician's role as agent.

Summary

- While there is no unambiguous definition of 'quality' in (primary) health care, most concepts embrace a technological aspect, e.g. best practice based on evidence, as well as individual and societal preferences.
- As a useful basis for economic analysis, a measure of quality could be employed that relates to a patient's health gain amended, perhaps, by a measure of the convenience attributes of care.
- Donabedian's distinction between 'structure', 'process' and 'outcome' aspects of quality provides a useful framework. A physician produces health outcomes using a stock of medical equipment and skills (structure) in a process that combines effort with a range of variable inputs including secondary care and pharmaceuticals.
- The physician's quality incentives relate to investment in skills and equipment (structure), as well as to the effort expended and the input mix chosen in the process of care. Quality incentives are shaped by the institutions of primary care and are subject to resource constraints. Institutions include the payment scheme, performance standards and the GP practice environment. Policy-makers shape the quality incentives by designing institutions.
- Policy-making aims at simultaneously achieving micro-efficiency (maximisation of a patient's or population's health gain given a set of resources) and macro-efficiency (optimising the level of resources allocated to (primary) health care) as well as equity. Conflicts in these aims usually lead to trade-offs between guaranteeing a high quality service and cost-containment, and between efficiency and equity concerns.
- Physicians act as agents on behalf of both the patient and the payer. This dual role has implications for their incentives and for the policies that shape them.

17 On the issue of equity in primary care provision see Gravelle and Sutton (2001).

3 THE PHYSICIAN'S OBJECTIVES AND QUALITY INCENTIVES: AN OVERVIEW

Much of the economic literature on physician services agrees on the fact that physicians maximise a utility function which includes income (or profit), a measure of patients' utility and sometimes social status (for an overview, see Scott 2000, section 4). Far less agreement exists as to the relative weights to be attributed to these arguments, although the US literature focuses most on the profit component (e.g. Dranove and Satterthwaite 2000; McGuire 2000).

A physician's utility can be understood as a function of monetary income (payment less monetary costs), professional and social status; the intrinsic benefit received from working; and, as a negative element, the non-monetary cost of effort (time spent). Status is attained as a social reward for an achievement in accord with a social norm. While such a norm can apply to a variety of merits such as medical success or income, status matters as an incentive only if the group amongst which it is achieved is sufficiently large and of sufficient concern to the individual. In contrast, an intrinsic benefit arises if the individual behaves according to their own values and norms, irrespective of the social circumstances. While intrinsic motivation is purely self-referential (a job well done), altruism takes into explicit account the patient's benefit.

With the GP seeking to maximise utility, incentives for the provision of quality are attached, in principle, to all components of that utility. Furthermore, it is possible to classify incentives according to whether the stimulus arises from competition, from regulation, or from values and social norms. Competition implies that the individual physician does not act in isolation but rather 'competes' with a number of rivals. In an economic context, competition is usually to win patients as a source of income. However, if status is an important source of utility, competition may also arise for status for its own sake. Regulation is usually employed as a stimulus if competition is too weak or if it gives rise to dysfunctional incentives. As with other professions, private or social values and norms are a powerful determinant of physician behaviour, and may even outstrip the role of competition or regulation.

Table 3.1 presents a framework for considering GP's incentives. It combines the sources of incentives (rows) with the physician's objectives (columns) on which they bear. The relationships between the various sources of incentives and the physician's utility are manifold and complicated. Any particular source usually has a bearing on more than one component of utility. Moreover, these relationships can be positive or negative, direct or indirect and frequently allow for feedback. The following examples corresponding to the boxes of Table 3.1 illustrates this complex web of incentives.

Competition gives rise to income related incentives if a GPs' remuneration increases with the demand for their services and if this demand is responsive to quality. This, in turn, depends on whether or not patients can somehow measure or experience the quality of a service and whether or not they can obtain the same or a substitute service from more than one provider.

Regulation has the most direct bearing on income related incentives by way of the payment system. The mode of remuneration crucially determines in which way, if at all, a physician's income is related to the quality of service. For example, a capitation payment per patient may induce a physician to use quality as an instrument to attract patients. However, if patient demand is not reactive to quality, capitation may induce the physician to cut back on quality in order to save cost. Sometimes payments are directly linked to measures of quality, e.g. in the forms of target payments or of fines to be paid for underachievement. Finally, regulation has an impact on a physician's income if failure to meet certain practice requirements leads to non-award or withdrawal of the medical licence.

While not being directly related to income, social norms can play an indirect role in that they shape how and to what extent competition and regulation act as sources of incentives. For example, if good professional behaviour constitutes an important merit in the view of society, then a good reputation is likely to carry a strong weight as one instrument in attracting patients. Professional norms bear heavily on the effectiveness of regulation. If good practice constitutes an important source of status, then self-regulation is likely to be effective. In contrast, if status is determined by income, the payment system and

Table 3.1 Classification of incentives

Physician's objective Source of incentive	Income	Status	Intrinsic benefit and altruism
Competition	Demand response, quality competition	Status competition in income or performance	Crowding out
Regulation	Payment system, performance indicators, fines, clinical governance	Published performance indicators, peer review	Crowding out, intrinsic benefit may depend on professional autonomy
Values and norms	Reputational rents professional or societal norm	Reputation relative to internalised norms	Work ethic,

other forms of income-related regulation are likely to play a more important role. Professional norms may also render regulation less effective if they foster a spirit of comradeship against external intervention.

I have already mentioned that physicians may compete not only for income but also for social status. An unstained professional track record may, thus, not only be instrumental in attracting patient demand but also something to vie for in its own right. Status competition is stronger the closer is the reference group, because an individual's behaviour is more easily observed within a close-knit group and because the award or withdrawal of social status is more immediate. Thus, one may expect professional status to play a stronger role within group practice than within the profession as a whole.

The effect of regulation on status as an incentive is ambivalent. On the one hand, the publication of performance indicators or the implementation of peer review is likely to enhance status competition by making an individual's performance common knowledge. On the other hand, regulation may erode a social norm and thereby devalue status as a source of motivation. This is the case if regulation lures members of society into the belief that good professional behaviour is being taken care of by the regulator and so does not involve an effort on the part of the individual professional that is worthy of social merit.

Even if professionals do not directly compete for status, social rewards may still serve as an incentive if the physician can improve social status amongst the population as a whole. Clearly, this requires a social norm somewhat wider than a purely professional norm.

It is sometimes argued that individuals are intrinsically motivated, in particular when carrying out tasks involving complex intellectual or manual activities or a certain amount of creativity. It is the feeling of doing a job well that motivates the individual, irrespective of the financial or social reward. A similar incentive arises for altruistic physicians who derive a benefit from improving the wellbeing of their patients. While these forms of motivation obviously require the presence of some private norms and values, they are also indirectly affected by the presence of competition and regulation. In particular, the presence of external incentives may crowd out intrinsic

motivation. This is the case if the individual practitioner perceives the provision of quality no longer as the product of their own initiative and effort but rather as something that has been forced upon them by competitive pressure or by the regulator's rewards. If regulation involves an increase in guidance and control, this may erode motivation, as the physician suffers a loss in autonomy.

It is worthwhile to note that important inter-relationships exist between some of the dimensions of incentives, which cannot be captured in a two dimensional matrix such as Table 3.1. For example, income and status are prone to interact. While status is likely to increase with income, a physician's income may also depend on professional status. There is also a link between regulation and competition. It is often an explicit aim of regulation to correct incentives that arise or fail to arise under competition. Indeed, some regulation is aimed at inducing (or mitigating) competition itself. Moreover, competition between physicians is important in those regulatory schemes that are based on relative performance. Finally, the profession's (rather than the individual physician's) status within society may determine the likelihood of regulatory intervention.

In the following five chapters, I discuss the literature concerning the different forms of incentives. As will be seen, the attention they have received varies widely. The bulk of the literature deals with economic incentives as they arise either from the demand response mechanism and competition (chapter 4) or from regulation, including the payment system (chapter 5) and more direct incentive pay or performance standards (chapter 6). Chapter 7 reviews the relatively scarce literature on status and intrinsic benefits as sources of incentives and comments on their interaction with market incentives and regulation. Chapter 8 then deals with clinical governance as a regulatory framework, which leaves some scope for non-economic incentives.

Summary

- Physicians can be viewed as maximising utility functions that contain in varying proportions: income; professional and social status; intrinsic benefits and altruistic concerns; as well as the cost of effort.
- Quality incentives can be attached to each of the elements of the utility function. The sources of these incentives are competition, regulation, the physician's ethical values, and social and professional norms.

4 INCOME RELATED INCENTIVES: DEMAND RESPONSE AND COMPETITION

Economists argue that the providers of a good or service choose quality and price in an optimal way only if there is a sufficient degree of competition in the market. For a number of reasons, this idea of competition does not carry over well to health care. First, insurance or state funding of services usually insulates patients against the cost of receiving care. Thus, one should expect them to be largely uninterested in the price charged by a physician. Second, physicians' fees or incomes are either regulated or are determined under corporate bargaining. In either case, they are taken as given by individual physicians. Third, a lack of patient choice between providers and/or imperfect information about the quality of care they offer may rule out effective competition.

In the following, I will outline how quality incentives could arise even if GPs' fees are determined centrally and covered by insurance. I then go on to address the conditions under which quality competition can arise and how it depends on the type of payment system. Deferring the discussion of non-pecuniary incentives until chapter 7, we assume for this chapter and chapters 5 and 6 that physicians seek to maximise their income. Income is given by the difference between revenue and the monetary cost of providing care. The variable cost of care is a function of the various inputs and the number of cases treated. Additionally, the physician usually bears a quantity-independent cost relating to investments in medical technology and skills. GPs' decisions on the amount of investment and effort to make, depend on a trade-off at the margin between the extra revenue generated and the extra cost. Thus the reimbursement system has important implications for the quality of care provided.

In the remainder of this chapter I analyse the conditions under which quality incentives arise in the presence of patient choice, which renders the demand for a GP's service sensitive to quality. In chapter 5 I use this framework to study the impact on quality of a range of payment systems (practice allowance, salary, capitation, FFS). It should be noted that although these payments bear on the physicians' quality incentives, they are not specifically linked to measures of quality. Aspects of such quality-related performance pay are the subject of chapter 6.

4.1 Demand response to quality

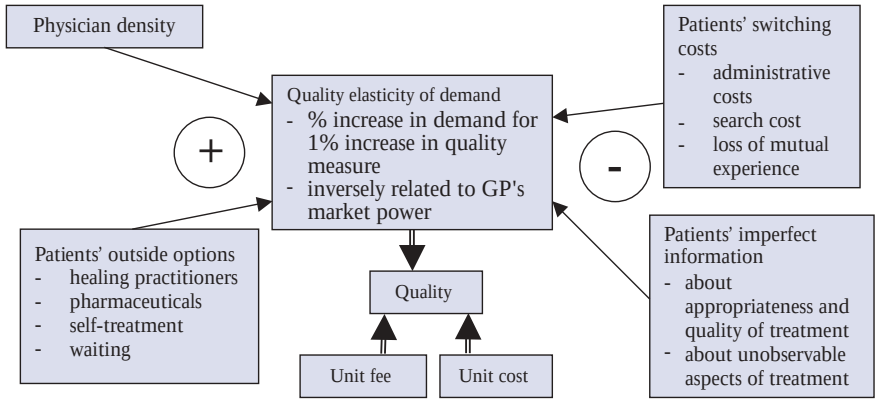
While most economic theory expresses demand as a function of price, it is clear that other determinants such as quality or advertising also have an impact. If patients face a travel or time cost or a non-monetary disutility from receiving care (e.g. side effects or physical or psychological discomfort), then their demand for a medical service increases with the quality of that care. Thus, even if patients receive care at a zero price, their demand for it can be expressed as an increasing function of the quality of the service (Ma 1994; Gravelle 1999; Chalkley and Malcomson 2000).[18]

When determining the quality of service by choosing the quality and quantity of inputs to use, including own effort and investment in medical skills, a physician has to make the following trade-off. If the fee rate is given, physicians may increase income by using quality to attract additional demand for their services. But this incentive is offset by the higher cost to the GP of providing a service of greater quality. Quality is then chosen at a level that will balance the increase in revenue from attracting additional demand with the increase in cost.

The extent to which the demand response mechanism stimulates the provision of quality depends on the quality elasticity of demand, which is defined as the percentage increase in demand for a one-percent increase in (a measure of) quality. A lower quality elasticity implies that demand is less responsive to quality and, thus, changes in quality have a weaker impact on revenue. Hence, the lower is the quality elasticity of demand the more reluctant the physician will be to provide quality for any given level of fee. Figure 4.1 illustrates the factors that influence the quality elasticity of demand, which, for a given fee and cost per unit, determines the quality of care offered.

18 Dranove and Satterthwaite (2000) consider a more general model, in which the demand for a medical service is a function of price, quality and the level of amenities.

Figure 4.1 Quality elasticity of demand and its determinants



As long as the quality elasticity of demand is not zero, a regulator/payer is able to stimulate the provision of quality by adjusting the fee it pays for primary care services. A higher fee per patient or treatment raises the physician's revenue generated from each unit of service. The GP is therefore willing to cater to more patients and raises service quality to attract them. The greater the quality elasticity of demand, the greater the quality stimulus from a fee increase. It follows that the level of fee at which the regulator can induce a certain level of quality is lower the greater is the quality elasticity of demand. If the elasticity is low, there is a strong trade-off between the provision of quality and financial feasibility.

In order to derive policy options to mitigate this problem, it is necessary to understand the determinants of the quality elasticity of demand. Some of the more salient factors (see Figure 4.1) are: the patients' other, or 'outside', options for care; the number of providers or physician density, i.e. the degree of competition; the information available to patients; and the costs they would incur in switching providers.

4.2 Quality competition in primary care markets

The decision of a patient to forego a GP's services depends on the other options that are available to deal with the health problem. These options may include self-treatment or consultation with an alternative provider of primary care. The more outside options that are available to a patient and the greater their effectiveness in solving the health problem, i.e. the better substitutes they are for the GP's care, the greater the quality elasticity of demand will be. The elasticity of demand (whether quality elasticity or price elasticity) is a negative measure of a provider's market power (e.g. Tirole 1988, chapter 1; Cabral 2000, chapters 5 and 9; Dranove and Satterthwaite 2000). The availability to the patient of alternative ways of improving health reduces a GP's market power by raising the degree of competition. In so doing, it enhances the provision of quality.

The provision of information and consultation (e.g. via internet or telephone) relating to self-treatment, the retailing of pharmaceuticals

and health care equipment for domestic use, and the supply of nursing or alternative medical services, all imply an increasing degree of 'outside' competition for GPs. The increasing provision or promotion of such services by GPs themselves could in some instances be interpreted as a reaction to greater outside competition.

Policy implication: One policy approach towards improving the quality of primary care lies in enhancing the availability of outside options to patients and patients' awareness of them.

Example: The NHS has recently undertaken considerable investment in providing patient information and better access to primary care (NHS Direct via telephone, online and from information points in key public places; walk-in centres staffed by nurse practitioners rather than GPs) (Department of Health 2000a).

One would expect the quality elasticity of demand for an individual GP's services to increase with the number of GPs. Gravelle (1999) considers quality competition between a number of GPs in a spatial set up, in which patients have to travel to receive a GP's services. Under competition, the quality provided increases both with the fee rate paid to GPs and with the number of GPs. Furthermore, quality decreases with the patients' travel cost. High travel costs imply that patients perceive different GPs' services as poor substitutes. Patients' reluctance to switch away from their local GP raises this GP's market power and, thus, weakens quality incentives.

A more difficult question concerns entry into the primary care market. Recent research in industrial organisation has shown that explaining the degree of competition by the number of firms in the market does not tell the full story (Cabral 2000, chapters 9 and 14). This is because the number of firms is not exogenous but rather is determined by the structure of demand and costs as well as by the form of competition. Generally, there is less entry into an industry where firms have to make initial investments that are high relative to the market size. For the primary health care 'industry', investments have to be undertaken in medical knowledge, premises and

equipment. One would expect that the greater these investments have to be, the lower the number of GPs within an area and, thus, the lower the competitive stimulus for quality provision. If an insufficient density of practices gives rise to a quality and access problem one could advocate subsidies on initial investments as one policy to stimulate entry.

However, from a social welfare point of view entry may be excessive rather than insufficient. This surprising possibility follows from the so called 'business stealing' argument. Firms enter as long as the expected operating profit covers their initial investment. The problem is that a substantial share of this profit is not due to an increase in social surplus but merely stems from 'stealing' rivals' business. This redistribution of profit is irrelevant for total social welfare. Therefore, firms may enter even if the cost of the investment exceeds the social gains from greater competition. Gravelle (1999) analyses the entry of GPs into the primary care market and shows that the fee rate under which GPs choose an optimal quality induces excessive entry. If the regulator/payer can only control the fee rate, a trade-off arises between financial efficiency and the provision of greater quality and access to health care.[19]

Recent research has pointed out that not only may the number of firms entering an industry be endogenous, but so too may be entry costs themselves (Cabral 2000, chapter 14). In the context of primary care, this implies that the level of (initial) investment in practice premises or medical skills is not only a determinant of entry and competition, but is determined itself by GPs' expectations about the degree of competition they are going to face. It follows that investment and entry have to be explained jointly by some underlying characteristics of the market, such as the size and social structure of the local catchment population, as well as by technology. A similar

19 Gravelle et al. (2002a) study the effect of regional attractiveness on GPs' location choices. GPs are willing to locate in less attractive areas only if they are compensated by a higher income. This requires them to have a larger patient list, given the UK's capitation-based GP payment system, which in turn implies a lower GP/patient ratio. Hence, unattractive areas are underdoctored, and primary care services are of lower quality in the sense of providing poorer access. Clearly, this raises equity concerns. The authors use the framework to study the effect of entry restrictions and other policies.

argument applies to the location choice of practitioners. These issues and their implications for the quality of, and level of access to, care remain to be explored.

4.3 Asymmetric information and patient choice

The extent to which competition can provide a stimulus for the provision of quality in primary care is likely to be severely constrained. In particular, incomplete information and switching costs on the part of the patients are prone to stifle competition and so weaken the incentives for quality enhancements. Let us consider the issues in turn.

The market for health care is plagued by the patients' imperfect medical knowledge with regard to diagnosis and therapy (Arrow 1963). The patient, therefore, has to rely on the physician as an agent, who is empowered to maintain or restore health. If physicians were to behave as perfect agents, i.e. in the best interest of their patients, then the informational problem would be resolved. However, it is not guaranteed that a particular physician possesses the skills required to serve a particular patient well. Nor is it guaranteed that the skills will be used to best effect. Lacking medical knowledge, patients are usually unable to evaluate key aspects of the service delivered. In order to evaluate a diagnosis and the appropriateness of the treatment proposed, patients need exactly the sort of medical knowledge the lack of which sends them to the physician in the first place. Hence, the agency relationship substitutes uncertainty about the physician (does the physician offer the right diagnosis and treatment?) for the patient's initial technological uncertainty (what diagnosis and what treatment?).

4.3.1 Hidden knowledge and investment in medical skills

The uncertainty the patient faces about the physician comprises two elements: 'hidden knowledge' and 'hidden action' (Hirshleifer and Riley 1992). 'Hidden knowledge' means that the patient is

20 Hence, the problem of hidden information relates to 'structure' aspect of quality as in Figure 2.1.

uninformed about the physician's skills in providing medical services.[20] A priori, only physicians are informed about their personal medical knowledge and skills and whether they are applicable to a patient's condition. Suppose now, that patients have a way of learning about a GP's skills. In this case, it is reasonable to expect that more patients seek to consult a skilled rather than an unskilled GP. In that case, there is a return to investments in medical skills by GPs if having more patients leads to a higher income for the doctor. By way of contrast, suppose that there is no way for patients to learn about a practitioner's skills and the quality of service. In this case, skilled physicians attract the same demand and income as unskilled ones and so do not receive a return on their extra investment in skills. If physicians are motivated by financial returns, then in this latter case investment in skills will be too low and quality will be under-provided: an 'adverse selection' has taken place.

Policy implication: The danger of under-investment in medical skills due to the potential inability of GPs to appropriate the returns to this investment justifies medical accreditation procedures on the basis of a compulsory curriculum as a way of ensuring at least minimum investment in knowledge and skills. Continuing professional education and periodic revalidation of licences are also aimed at safeguarding quality levels (Department of Health, 2001 a,b).

An issue arises about who should bear the cost of this 'forced' investment. If accreditation is based on the argument that physicians would not otherwise undertake the necessary investments, this reveals that these investments are unprofitable. If accreditation enforces such investment and if the cost has to be born by the (prospective) physician, it is possible that this deters those individuals from entering the profession who have the option of following a career which generates higher returns to their investment. This, in turn, would provide a case for subsidy by the policy-maker of compulsory investments in medical skills.

4.3.2 Hidden action and moral hazard

In the case of 'hidden action' the patient cannot infer the GP's effort in delivering a service of appropriate quality.[21] Here, the incentive to shirk on quality, a 'moral hazard', arises as the physician tries to economise on monetary or non-monetary cost. Similarly, the patient may be unable to judge whether the mix and level of services suggested by the GP is appropriate. Again, the informational asymmetry is likely to entail an under-provision of quality.[22]

More generally, the quality elasticity of demand increases with the level of patient information (Dranove and Satterthwaite 1992, 2000). When poorly informed about quality, patients rationally reduce the weight they attach to quality. But in that case a patient becomes less sensitive to quality changes. The associated reduction in the quality elasticity of demand implies a lower incentive for physicians to provide quality.

As Chalkley and Malcomson (1998a, 2000) and Dranove and Satterthwaite (2000) point out, patients are usually able to observe some quality attributes better than others. For instance, they may be good judges of the amenities of a GP's practice and the friendliness of the GP and other practice staff but poor judges of medical expertise. From a GP's point of view, there is then an individual elasticity of demand attached to each dimension of quality, with the elasticity with respect to amenities being greater than the elasticity with respect to medical expertise. The physician will choose each dimension of quality individually with a resulting bias towards those dimensions that are easily observable and to which patient demand is consequently more responsive. Specifically, this implies a bias towards convenience attributes of care or towards treatments that promise a short-term success. Furthermore, practitioners may shy away from recommendations to the patient that yield health gains eventually but bring with them a short-term disutility.

21 Hence, the problem of hidden action relates to the process aspect of quality.

22 Here, I am considering 'adverse selection' and 'moral hazard' on the supply side. The same concepts are also applied to the demand for insurance (e.g. Zweifel and Breyer 1997, chapter 6), an issue I do not address here.

4.3.3 Search, signalling and reputation as solutions to informational asymmetries

Industrial organisation theory suggests a number of mechanisms by which information about quality can credibly be communicated to patients such that quality incentives are restored. These approaches include patient search, signalling, reputation and regulatory measures.

Rochaix (1989) models a health care market in which patients engage in search by consulting a number of physicians, each of whom propose a treatment. Self-interested physicians would have an interest in distorting the intensity and structure of the treatment in a way that maximises their income even if at the expense of the patient. By way of simulation, Rochaix (1989) shows that an increase in search cost or in the urgency of the patient's need tends to exacerbate this problem. In contrast, a wider distribution of medical information amongst patients reduces the problem. Even the presence of a few well-informed patients entails a benefit for the whole patient population.[23]

Policy implication: A reduction in patients' search costs provides one rationale for the provision of public information on physicians and their performance. Performance indicators and some of the problems related to them will be discussed in section 4.4 below.

In many cases, however, primary care services may be viewed as an 'experience good'. The patient can truly evaluate the quality of a GP's services only after having experienced them. The initial choice of GP may still not be arbitrary but may rather be based upon signals or reputation. Unfortunately, the literature has so far not brought forth signalling or reputation models that are applicable to European health care systems, in which physicians are unable to control their fees. However, despite being founded on a framework involving market prices, the models by Shapiro (1986) and Biglaiser and Friedman

23 If patients could invest in medical knowledge in order to reduce their risk of being under-treated, there would still be under-investment, from a societal point of view, in acquiring medical knowledge. This is because, when investing, individuals do not take into account the external benefit they bestow upon less informed patients.

(1994) give some insights about the mechanisms underlying signalling and reputation.[24]

Signalling and reputation are closely linked to the problems of 'hidden information' and 'hidden action'. If quality is determined by the physician's skills, a patient who is unable to observe them faces a problem of hidden information upon the first consultation. Quality being an experience attribute of the service, the patient is able to infer the GP's skills ex-post and is informed thereafter. In this regard, asymmetric information only lasts for a single period. However, skilled providers still face the problem of signalling their ability to prospective patients in order to guarantee that their service is selected from the outset.

Credible signalling involves the physician undertaking a costly activity, which can be observed by patients and which only a skilled GP would engage in. Some such activities could be: advertising; the offer of extra services, e.g. longer consultation hours or home-visits; investments in practice premises and medical equipment; or engagement in continuing medical education over and above the required level. Skilled physicians can then credibly signal their type by choosing a level of such activities that is still profitable to them but would be unprofitable for unskilled types.

Here, the possibility of repeat visits is important. Suppose a GP is able to signal skills to a patient. Upon having experienced high quality this patient will consult the same physician again. In contrast, even if a quack is able to attract a patient by mimicking the skilled GP's signal, the patient will not return after having experienced a low quality service. It follows that the value of gaining a patient is always higher for a skilled GP than for a quack.[25] A skilled GP could then increase their signalling activity to a level at which its cost exceeds its (lower) value to the quack. Since the latter would, therefore, not engage in this activity, this becomes a credible signal of skills. In this

24 For an excellent introduction to these issues, see Tirole (1988, sections 2.3-2.6).
25 This is the case even if costs are taken into account. When using the same amount of inputs, and so incurring the same cost, a skilled physician can always render a higher quality service. Thus, there is no cost advantage for an unskilled GP. This argument may break down, however, if skilled GPs face higher opportunity costs of time and effort.

regard, practice advertising, or the attainment of extra-qualifications on the part of GPs may be viewed as a form of quality signalling.[26] Note that skilled GPs incur the extra cost associated with signalling as the price of distinguishing themselves from less skilled rivals.

While experience may reveal to patients a GP's type and, thereby, resolve the problem of 'hidden information', the same does not apply to the problem of 'hidden action'. Current experience of a high effort by a provider does not allow a prediction about the provider's future effort. Signalling is, therefore, ineffective in resolving the problem of 'hidden action'. Here, the mechanism of reputation plays a role.

Suppose a GP has established a reputation for providing a high quality service.[27] Patients consulting this GP initially trust that they will continue to receive high quality care. The GP expects that if they were to reduce quality and thereby fail to honour that trust, then they would be punished by some patients who would switch to a different provider. The ensuing reduction in demand for the GP's services reduces future revenue. The GP has to trade-off the cost saving from a reduction in quality against the loss in future revenue. Thus, the desire to maintain a good reputation is more effective as a guarantor of quality the greater the difference between the fee per unit and the marginal cost for the high quality treatment and the greater the loss of demand that would follow a reduction in service quality.

Patients are only able to 'punish' the GP for cutting quality by switching away if they are able to perceive the reduction in quality. This requires some medical knowledge, as well as a sufficient rate of utilisation of the GP's services. Furthermore, the degree of response to a cut in quality by a provider is enhanced if patients communicate with one another, but is dampened if there is a strong inflow of inexperienced patients.

26 The idea of advertising as a signal follows Kihlstrom and Riordan (1984). Incidentally, advertising may be a signal of quality even if its content is entirely uninformative. A public 'burning of money' would be an equally informative signal. The idea of education as a signal follows Spence (1973).
27 The following argument is strongly based on the intuition of Biglaiser and Friedman (1994).

Policy implication: This underlines a case for supporting the formation of information fora that facilitate the dissemination and exchange of information amongst patients.

Finally, the GP's discount rate plays an important role in determining the weight that they put on the future loss in revenue. For instance, a GP who expects to retire soon will have little concern about a loss of future revenue and will, therefore, be more inclined to reduce the quality of care. Incidentally, this may be signified by some older GPs choosing not to keep up with medical progress by updating their skills and equipment.

Policy implication: Such arguments point to the value of continuing professional education and regular revalidation, which are the subject of recent policy in the UK (Department of Health 1998, 1999, 2001 a, b).

So far, we have not addressed how a physician can acquire a good reputation in the first place. Shapiro (1986) and Biglaiser and Friedman (1994) provide some insights by combining signalling and reputation models. One way of acquiring a reputation is to sell a high quality service 'under par', i.e. to provide high quality even if there is no return in terms of greater revenue. In so doing, providers may attract a high level of initial demand from patients with whom they establish a reputation for quality. A fast track to reputation may lie in sending a signal of high quality to begin with. If consumers believe this initial signal, they flock to the newcomer, and from this point on the reputation mechanism takes over. In this regard, signalling may be an attractive strategy for skilled GPs who have yet to establish a reputation.

A limitation of the mechanisms of search, signalling and reputation is that patients are imperfect in their judgement of the quality experienced. In particular, the inherent uncertainty of the success of medical treatment gives rise to a distribution of possible health outcomes, which is difficult for the patient to interpret. Health care is characterised by 'credence attributes', i.e. features that are open only to expert judgement. As argued earlier, lack of expertise on the

part of patients leads them to employ physicians as their agents. Given that, it would be unreasonable to presume that patients can fully evaluate what experts are doing (Arrow 1963). This problem is most pronounced in the context of specialist, i.e. secondary, care, whereas the patient may develop some capability for judging the quality of primary care. Indeed, Dranove and White (1987) explain long-term physician-patient relationships by the reduction in the patient's monitoring cost they enable – or, conversely, by the increase in patient expertise. This preserves some scope for signalling and reputation in ensuring quality.

If patient experience reflects an imperfect view of the quality of a practitioner, then the collective reputation of the GPs working in a group practice or, indeed, the collective reputation of the profession as a whole, may to some extent replace individual reputation.[28] Here, a collective reputation may be viewed as the overall impression of the profession or of a group of professionals that a patient gathers from their own experience, from word of mouth or from media coverage. For the professionals, collective reputation is a form of public good, where each physician's reputation is to some extent determined by the actions of their peers.[29] In such a case each agent has an incentive to free-ride on their peers and to under-provide effort in maintaining the profession's reputation.

Tirole (1996) considers an interesting variant to this theme where, in an inter-temporal framework, the reputation of an agent today is determined by the behaviour of past generations. One of Tirole's findings is that the bad reputation acquired by one generation of agents in the past can reduce the financial returns to a good reputation by so much that the incentives for succeeding generations to acquire it are destroyed. Thus, a bad reputation can be self-reinforcing over time. In that case only external intervention can restore quality incentives as well as the profession's reputation.

If collective reputation fails as a mechanism in maintaining quality, then the only way to guarantee the quality of care, given its credence

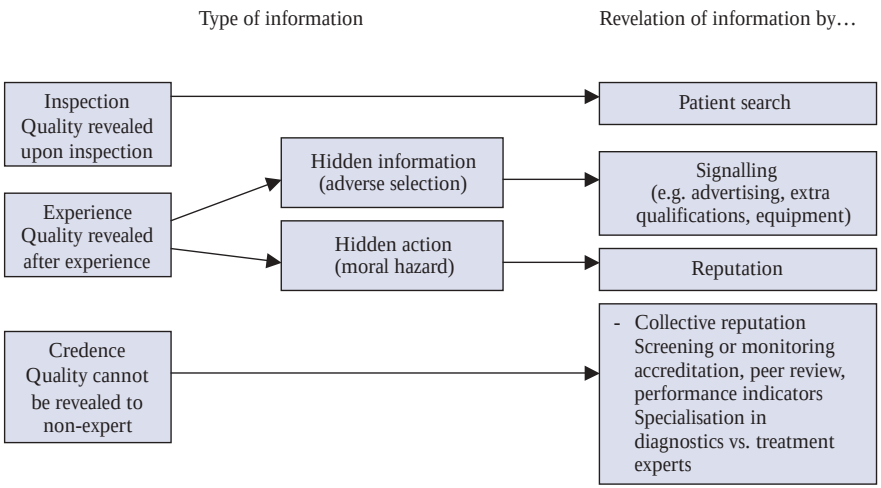
28 Collective reputation is particularly important when patients have to change physicians repeatedly.

29 For an application of this idea to a physician partnership, see Getzen (1984).

attributes, lies in the screening and monitoring of a physician by an expert auditor. The information could then either be used in direct regulation of the provider or it could be transmitted to patients in the form of performance indicators.

Figure 4.2 summarises the different forms of (incomplete) information and the possible mechanisms that may lead to the resolution of informational asymmetries.

Figure 4.2 Types of information and mechanisms of revelation



At this point let us note that all of the mechanisms discussed – search, signalling, and reputation – involve a welfare cost corresponding to the value of information. In the search scenario, this cost is borne by the patient who risks a deterioration of their condition and bears a time cost when seeking better provision. In the signalling case, the cost is borne by high quality providers who over-invest in advertising, skills acquisition or practice amenities. If the investments undertaken for signalling purposes raise the physician's skills or improve the available medical technology, then the welfare cost of information may be contained; yet there is an inefficiency relative to the case of perfect information. Finally, reputation gives rise to a welfare loss in that it requires the incentive of a fee well in excess of the marginal cost of high quality treatment.

4.4 Regulatory measures to reduce asymmetric information

Let us now turn to a number of regulatory measures, which support the resolution of the problem of asymmetric information. To begin with, note that the regulator can support both the signalling and the reputation mechanisms by raising the fee rate paid to GPs. This increases a GP's returns from the repeat purchases they obtain if they offer higher quality provision.

Regulatory instruments that tackle the informational problem in a more straightforward manner include certification, professional accreditation (Shapiro 1986) and the publication of performance indicators (Dranove and Satterthwaite 2000; Gravelle and Masiero 2000). Certification corresponds to a publication of the physician's investment in human capital. We have already addressed the way in which this may help a physician to signal superior skills.

Accreditation corresponds to an arrangement under which a GP can set up practice only when satisfying a minimum requirement regarding medical knowledge. Obviously, such a policy can help to screen out providers with the least satisfactory levels of skills. As discussed earlier, accreditation can thereby directly alleviate the 'hidden information' problem. Moreover, if the additional cost of

providing quality decreases with the level of skills possessed by the GP, then accreditation also supports reputation as a guarantor of quality and can thereby indirectly mitigate the 'moral hazard' problem.[30] Recall that the incentive to reduce quality increases with the potential cost savings from doing so. As highly skilled physicians find it easier to provide a higher quality, they gain less from cutting quality and will, therefore, be less inclined to do so when this entails a loss of reputation. Accreditation helps to screen out those poorly skilled candidate physicians for whom maintaining a reputation is the least likely to work as a guarantor of quality.

Example: A recent English NHS consultation paper on preventing, recognising and dealing with poor clinical performance by doctors stresses the role of accreditation, revalidation and credentialling for doctors and locums (Department of Health 1999). 'Credentialling' involves verification from doctors' performance records of whether they have been the subject of action by relevant regulatory bodies; of whether they were involved in a high number of complaints or litigation cases; and of whether they had actually obtained the qualifications listed in their CVs.

By publishing performance indicators, the regulator can improve patients' information (Dranove and Satterthwaite 2000, section 6.3). This in turn is expected to increase the quality elasticity of demand and to enhance quality competition. If the chosen performance indicators are positively correlated with true quality, this goal may be achieved. However, the design of such indicators is by no means straightforward (Blumenthal and Epstein 1996). Indicators may omit important but hard to measure dimensions of quality. If the provision of care amounts to more than the sum of its parts, it may also be misleading to measure quality of care by a set of individual indicators. Furthermore, the use of indicators may trigger a number of dysfunctional reactions on the part of physicians (Smith 1995). These range from outright manipulation to an undue focus on those

30 This (side) effect is present, irrespective of whether or not is has been taken into account in the design of the accreditation scheme.

dimensions of performance that are being publicised. Finally, patients may be unable to interpret the indicators for lack of understanding and, therefore, ignore or even misread them (Hibbard and Jewett 1997; Mennemeyer et al. 1997).

Giuffrida et al. (1999, 2000) show by way of econometric analysis how difficult it is to interpret hospital admission rates as quality indicators for primary care. They conclude that due to their high variability over time, indicators should only be used as moving averages. Furthermore, indicators should cover only those aspects of care – but all of them – which can be controlled by GPs. In an example they demonstrate how a ranking of health authorities with regard to the indicator 'rates of admission to hospital' vary when crude rates are step-by-step adjusted for age and sex; morbidity factors; socio-economic factors; and the supply of secondary care. For instance, Manchester holds the first rank (i.e. most admissions) with regard to the crude rate of admissions to hospital; drops to rank 9 when morbidity and socio-economic factors are accounted for; and drops out of the top ten when the supply of secondary care is also included. This underpins how demanding a task the development and use of quality indicators is.[31]

Example: In the UK NHS, National Service Frameworks specify for five key areas (Coronary Heart Disease (CHD), Diabetes, Older People, Mental Health and Cancer) national performance targets which are subsequently broken down to the level of primary or secondary care providers. For the relevance to primary care see Department of Health 2002a. Specific performance indicators are being developed. The example in Box 4.1 is taken from the National Service Framework for CHD (Department of Health 2000b).

31 For a further discussion of indicators and the difficulties in implementing them see McColl et al. (2000) and Wilkinson et al. (2000).

Box 4.1 Example of performance targets – UK National Service Framework for Coronary Heart Disease

Standard three: GPs and primary care teams should identify all people with established cardiovascular disease and offer them comprehensive advice and appropriate treatment to reduce their risks.

Standard four: GPs and primary health care teams should identify all people at significant risk of cardiovascular disease but who have not yet developed symptoms and offer them appropriate advice and treatment to reduce their risks.

From this derive the following 'milestones' for primary care practices.

- *Milestone 1: By October 2000 every practice should have: clinical teams that meet [...] at least once every quarter to plan and discuss the results of clinical audit and [...] clinical issues.*
- *Milestone 2: By April 2001 every practice should have: all medical records and hospital correspondence [...]; appropriate medical records containing easily discernible drug therapy lists for patients on long term therapy; a systematically developed and maintained practice-based CHD register [...] which is actively used to provide structured care [...].*
- *Milestone 3: By April 2002 every practice should have: a protocol describing the systematic assessment, treatment and follow-up of people with suspected angina [...] is being used to provide structured care [...].*
- *Milestone 4: By April 2003 every practice should have: clinical audit data no more than 12 months old [...].*

These milestones relate to the following NSF goal.

NSF goal: Every practice should: deliver or offer advice about each of the specified effective interventions to all of those in whom they are indicated, demonstrated by clinical audit data no more than 12 months old.

Performance is measured or will be measured according to the following indicators.

Health improvement

- *Age standardised or age and sex standardised CHD mortality rates by Health Authority (and 10 yearly, by socio-economic class).*

Fair access and effective delivery of appropriate health care

- *The number and % of practices in a [primary care organisation] with a systematic approach to following up people with CHD (new collection from 2001/02).*
- *The number and proportion of people aged 35 to 74 years with recognised CHD whose records document advice about use of aspirin.*

Patient/carer experience of NHS

- *The national survey of CHD patients will provide information on variations and provide a baseline against which future surveys will be analysed.*

Health outcomes of NHS care

- *Age-sex standardised rate of cardiovascular events in people with a prior diagnosis of CHD, PVD, TIA or occlusive stroke.*

Source: Department of Health 2000b.

While these indicators apply at the level of primary care organisations containing many GP practices, it is expected from them that they implement, monitor and enforce similar performance standards at individual practice level (Department of Health 1999).

Dranove et al. (2002) provide an interesting analysis of how complicated the effects of imperfect performance indicators are. The use of death rates as performance indicators for heart surgeons by a number of US states has prompted a great deal of criticism, as these rates are only imperfectly adjusted for case mix. They are open to manipulation since surgeons can turn down severe cases and thereby

reduce their risk of being indicated as poor performers. In the presence of some, albeit imperfect, risk adjustment, skilled physicians are more inclined to accept severe cases. However, the indicators may still induce the shunning of high-risk patients across the board.

Furthermore, they are likely to induce a shift in practice towards less risky procedures, such as angioplasty. Dranove et al. (2002) argue that the latter two effects may be welfare improving or reducing depending on the impact both on health care costs and health outcomes. An empirical analysis carried out with data for the states of New York and Pennsylvania showed that the net effect of report cards was an increase in health care resource use and poorer health outcomes. The more general lessons from this analysis extend to the less contentious context of primary care and demonstrate the complexity of the relationship between indicators and physician behaviour as well as the scope for unintended effects.

I have argued above that the resolution of asymmetric information generally leads to a welfare loss, which can be interpreted as an indirect cost of quality provision. Consequently, when there is asymmetric information, a regulator should seek to impose a lower level of quality than would be optimal under full information.

4.5 Patient switching

It should be expected that, if they change GP, patients incur substantial switching costs in the form of the foregone benefits of a long-term physician-patient relationship. As Dranove and White (1987) emphasise, such a relationship is mutually beneficial in reducing the patient's monitoring cost and the physician's cost of diagnosis, where both types of cost are negatively linked to the quality of care. Thus, patients should be expected to have a disincentive to switch physicians. This, however, reduces the patient's quality elasticity of demand. As soon as a patient is locked into a long-term relationship with a physician, the latter may have an incentive to lower quality.

Gravelle and Masiero (2000) show that this need not always be the case. They consider a model of quality competition between two GPs

and across two time periods.[32] Patients switch practice after the first period if perceived quality falls short of their expectation. As patients' perceptions are subject to measurement error, some erroneous switching takes place. Gravelle and Masiero (2000) show that the provision of quality is independent of the switching cost. Furthermore, although some patients switch erroneously, the ability to switch is socially desirable.

Finally, the presence of patient error does not inhibit the ability of a regulator/payer to use capitation fees as an instrument to establish a socially optimal level of quality. However, if patient error increases with the level of quality – implying that predicting reliably that a doctor is good is more difficult than predicting reliably that a doctor is poor – the volume of erroneous switching increases with the level of quality. The associated cost of switching constitutes an indirect cost of quality and the regulator, therefore, chooses a level of quality under imperfect information that is below the optimum level under perfect information.

4.6 Empirical evidence on patient choice

Little evidence is available as to the presence of quality competition within a European primary care context.[33] Two strands of literature relate to patient choice of GPs or primary care practices (reviewed in Scott 2000, section 3). The first set of research analyses patients' stated preferences over a range of quality attributes relating to the physician-patient relationship or to practices. A problem of most of these studies is that by focusing on general survey data or data relating to an individual practice they ignore the aspect of patient choice between practices.

32 The assumption of a duopoly is a legitimate simplification if patients have to incur travel costs to reach GPs. The model then embraces the catchment area of patients located between the two GPs. The results are easily generalised when considering a great number of these localised markets.

33 Dranove and Satterthwaite (2000) report some evidence on the extent of quality competition in the US hospital market.

This has been addressed by Vick and Scott (1998), who use a discrete choice experiment to evaluate the relative importance of factors such as waiting times or attributes of the GP-patient relationship. A bottom line to this as well as to previous studies is that patients tend to attach most weight to their relationship with the doctor (e.g. 'the doctor listens', 'being able to talk to the doctor'). This highlights the important role of the doctor-patient relationship and the need for communicating private information not only from the doctor to the patient but also from the patient to the doctor.

Variables relating to clinical quality have usually not been included in the models. However, Vick and Scott (1998) find that patients attach least importance to being involved in treatment choice ('who chooses your treatment'). The low preference for involvement suggests that patients are not comfortable about assessing the clinical quality of the treatment. Taken together with the strong preference for a good relationship, this lends some indirect support to the hypothesis that patients attach more weight to attributes that they are better able to assess (the relationship as opposed to the clinical performance of the doctor).

A second path of investigation is followed by Dixon et al. (1997). Using patient level data from three English health authorities, the authors examine the determinants of patients leaving GP practices other than because of having changed address.[34] Such behaviour may be interpreted either as patients leaving their prior practice due to discontent or as patients switching to a practice that offers them a superior service.35 Econometric estimations suggest the following.

First, that patients are less likely to leave practices that have more GPs, or offer more clinics or longer opening hours. This may indicate that at least some observable aspects of quality play a role in patient choice. Older patients are less likely to switch practice, which may be due to greater switching costs within a long-term patient-physician relationship. Alternatively, one could hypothesise that older patients

34 Until April 2002, Health Authorities were nearly 100 geographically delineated NHS agencies responsible for planning and securing the provision of health care for local populations (average 0.5 million). From April 2002, planning and commissioning responsibilities transferred to over 300 Primary Care Trusts.

have, by way of previous switching, found a physician to their satisfaction. Female patients are more likely to abandon a GP practice. It remains open to interpretation whether this relates to a greater sensitivity in inter-personal relationships, or whether women are inherently more concerned about quality. Patients are also more likely to leave a practice if they live in deprived areas. Whether or not this implies that patients from such areas are more responsive to quality differentials remains unresolved. On the one hand, one might expect the average level of patient education to be lower in deprived areas, a fact that would hint at lower quality sensitivity. On the other hand, the time cost of practice shopping, in terms of foregone wages, may be low in deprived areas, indicating higher quality sensitivity.

While these results suggest some scope for quality as a determinant of patient choice between GPs, the overall level of patient switching – other than when moving to a new address – is very low, at 1% to 1.5% of patients per year. Such low switching rates do not necessarily imply that quality competition is irrelevant, however. Firstly, in a situation of equilibrium, after GPs have adjusted their quality and patients have chosen their preferred practice, one should not expect a high volume of switching. Yet the provision of quality still matters in such a state. Only by continuing to provide quality can GPs maintain their patient base. Were individual GPs to lower quality they would lose patients. In this regard, the list size of a GP may be a better indicator of quality once controlling for other determinants. Unfortunately, data on list sizes were unavailable to Dixon et al. (1997).

Second, even if the low levels of switching are due to the existence of substantial switching costs, GPs still have an incentive to compete for patients in order to establish their practice in the first place. In this respect, the group of patients who switch due to a change of address becomes relevant, as these are the potential 'new' patients in a region for which GPs compete. Even if these patients lack experience, their choices may be guided by a GP's reputation for quality, as may be communicated to them by experienced patients. Unfortunately, Dixon et al. (1997) do not provide any evidence as to the determinants of patients' initial choice of practice.

4.7 Discrimination

We have so far put to one side the fact that patients vary in their characteristics, such as the severity of their illness, their susceptibility to treatment and the state of their medical knowledge. But it is reasonable to presume that GPs are able to determine at least some of the above patient characteristics by way of diagnostic consultation, particularly during a long-term relationship. That implies scope for non-altruistic physicians to devise ways in which to discriminate between patients so as to maximise income (Allen and Gertler 1991).

Technically, patient heterogeneity is reflected in patient-specific quality elasticities of demand. For example, patients suffering from a condition that warrants urgent attention, patients who face high search costs, or patients who are poorly informed tend to exhibit a lower quality elasticity of demand. Profit-maximising GPs who receive a uniform payment per patient would offer lower quality to those patients with a lower quality elasticity of demand. In so doing, they can rest assured that these patients will still not switch away. Conversely, patients who are well informed about their illness and who can afford to shop around may receive higher levels of quality. There is an obvious equity issue as patients are likely to receive care not according to their need but rather according to their information or patience. The pattern of inequality depends strongly on the nature of patient heterogeneity, where clear-cut predictions are difficult to make. General policy implications are, therefore, not easily derived.

Allen and Gertler (1991) show that the issue of quality discrimination does not only touch on equity but also on the efficiency of service provision. If the regulator is constrained to set a uniform capitation payment for all patients, which is a realistic assumption in the context of primary care, a monopolistic provider tends to distort quality away from the patient specific welfare-maximising levels. If, for example, the capitation rate is set according to an average of the treatment costs for severely and less-severely ill

35 Patients who have switched practice with a change in address are in most cases likely to have done so to find a practitioner who is located closer to their new address and not for reasons related to the quality of service.

patients, a situation arises in which the quality supplied to severe (less-severe) cases falls short of (exceeds) the optimal levels. This is in order to deter (attract) patients for which treatment cost is above (below) average. This form of under- or over-provision of quality is sometimes referred to as skimping and cream-skimming respectively.

Ellis (1998) confirms these findings for a duopoly and shows in addition, under which conditions high cost patients are being dumped, i.e. struck from physicians' lists.[36] An alternative practice is refusal to admit bad prospects to the list. Indeed, this latter form of discrimination is more likely in that it is easier to carry out and easier to reconcile with professional ethics, at least superficially.

In order to avoid the various forms of discrimination, a regulator would need to set patient specific-capitation payments, which could induce GPs to provide appropriate quality for each patient. But such a policy is not feasible, due to informational constraints on the part of the regulator and to the large administration costs involved.

Policy implication: The regulator has to rely on imposing minimum quality standards in order to avoid the worst forms of skimping or dumping. Furthermore, there is a case for targeting policies at patients with particularly low quality elasticity of demand. If, for instance, quality discrimination occurs according to the state of patient information, then this provides a good justification for information campaigns being targeted at the poorly educated.

Dixon et al. (1997) studied whether patient discrimination plays a role for practices in the three English health authorities they surveyed. They concentrated their analysis on discrimination in the form of admission to a GP's list. The very low average of 0.2% patients per year being removed from lists at the GP's request confirms that, if at all, discrimination takes place mainly in the form of non-admission. Dixon et al. hypothesise that the incentive to cream-skim patients is greater for fundholding practices (see section 5.3 below) as these are

36 There are, of course, justifiable reasons, such as violent or abusive behaviour, for patients being struck from a physician's list.

not only exposed to the costs of primary care treatment but also to the cost of secondary care referrals.

But this hypothesis is only partially confirmed. Lacking data, they only consider admission to GP lists of elderly patients (over 65). Here, they find that the proportion of elderly patients transferring into a practice is greater for non-fundholding practices but only if these patients have moved from outside into the (health authority's) region. Fundholding practices received a budget for the commissioning of some secondary care services and for the purchase of pharmaceuticals. One might conjecture that they had an incentive to contain the number of elderly – and presumably more costly patients – on their list. This would, indeed, show up in the data as a more than proportionate transfer of elderly patients into non-fundholders who were not directly exposed to the cost of care. Dixon et al. (1997) explain the finding that patients transferring within a region did not seem to have been subject to cream-skimming by referring to their superior information on local practices which puts them in a better position to dispute a refusal.[37]

37 According to Campbell et al. (2001a), Black, South Asian and Chinese respondents, as well as respondents from less affluent groups, reported systematically lower satisfaction in a survey based on the General Practice Assessment Survey instrument. It remains unclear whether this can be interpreted as a result of discrimination. The differences in reported satisfaction rates are likely to be affected by geographical correlation: GPs in deprived areas, which may be characterised by a higher share of non-white population, may face a more difficult case mix and 'harsher' working conditions across the board. If the poorer outcomes are reflected in lower satisfaction scores, the differences in scores are reflective of geographical inequalities rather than of discrimination.

Summary

- In as far as physicians maximise income, they trade-off at the margin the revenue generated from the provision of quality against the cost of doing so. The reimbursement system shapes the nature of this trade-off and is therefore instrumental in the provision of quality incentives.
- GPs' quality incentives increase with the margin between unit fee and unit cost and with the quality elasticity of demand, i.e. the percentage increase in patient demand for a one-percent increase in the quality measure.
- The quality elasticity of demand increases with physician density and the availability of alternative sources of care, and decreases the greater are patients' switching costs and the less information they have on the quality of the service.
- GPs are induced to increase the quality of their services if they compete against 'outside' offers (e.g. tele-medicine, self-treatment or alternative medical services) or against each other.
- Asymmetric information about the physician's skills (hidden information) and effort (hidden information) restrict patient choice and stifle competition. It is likely to lead to an under-provision of quality in general, or to the provision of quality only in dimensions that are observable to the patient, e.g. practice amenities.
- Asymmetric information can be resolved by the following mechanisms: search, if the patient can inspect the relevant quality aspect, e.g. the physician's equipment or skill-related certificates; signalling of hidden information by the physician (e.g. the acquisition of extra certificates) or the acquisition of a quality reputation if the service is of an experience nature; collective reputation or professional credentialling by independent experts if the service is of a credence nature so that its quality cannot be determined by patients.
- Regulatory measures to reduce informational asymmetries include (re-) accreditation, certification and the use of performance indicators. Generally, the resolution of asymmetric information is

costly from a social point of view. This cost has to be counted as an indirect cost of quality.

- Long-term relationships between GPs and their patients enhance the provision of quality in that they reduce the informational asymmetries between patient and doctor. However, they also raise the patient's cost of switching to a different GP and thereby reduce quality competition.
- Survey evidence confirms that patients are uncomfortable about assessing clinical aspects of quality and thereby provides indirect support for concerns that clinical quality may be under-provided. Empirical analysis of patient switching finds low levels of switching in general. However, this does not necessarily imply the absence of quality competition; as in equilibrium low levels of switching should be expected.
- If some patients exhibit a lower quality elasticity of demand, e.g. due to lack of information or because their need for treatment is urgent, physicians may be tempted to discriminate and under-provide quality to those patients. This leads to a loss of efficiency and equity.

5 PHYSICIAN REMUNERATION AND PROVISION OF QUALITY

In the previous chapter, I have discussed how the demand response mechanism and competition can induce physicians to provide quality. While I have presumed that physicians receive a fixed fee for their services, I have remained unspecific about the nature of this fee. This chapter considers in greater detail the various forms of physician payment and their implications for the provision of quality.

5.1 Fixed budgets

Zweifel and Breyer (1997, section 7.3) consider a GP who receives a fixed budget and produces two medical care services by combining use of time – in our terminology effort – and another input, which has to be purchased. The GP has to bear the full monetary and non-monetary costs of service provision to a number of patients. The authors find that while the GP provides an efficient mix of services, the levels of these services tend towards a minimum unless the GP is driven by non-pecuniary incentives. Even if demand increases with quality, this does not raise revenue, only cost.

When receiving a fixed budget, a non-altruistic GP has no incentive to provide services in excess of some minimum imposed by liability rules or by a regulatory standard. Likewise, there is no incentive for the GP to invest in skills or medical capital beyond the minimum required, as the financial return to additional investments is zero.

Zweifel and Breyer carried out their analysis in a setting with certainty. In reality, GPs face uncertainty, for example with regard to the number and severity of cases arriving over time. If they are averse to the risk of over-spending their budget, this is then likely to distort the structure of care over time and the mix of inputs. Glazer and Shmueli (1995) consider a situation in which patients arrive sequentially and in which the physician is uncertain about the number and severity of future cases. They show that under these circumstances, a global budget generally leads to an unequal distribution of services provided to patients with equal need if they arrive at different points in time. Here, the care provided to patients

arriving at a later date depends on the residual budget. The level of care provided to patients arriving early depends on the physician's risk attitude.

Thus, cautious physicians will save on early patients and under-provide services, whereas less cautious physicians may over-provide services to those patients. In either case, the conditions for horizontal equity (same volume of care provided to patients with same health) and vertical equity (greater volume of care provided to patients with poorer health) are usually violated.

The situation is even more complicated if a number of physicians compete for their share of a budget, as GP practices might within an English Primary Care Trust (Dusheiko et al. 2001) or under the global budgeting arrangements in some continental health care systems. In this case, an incentive exists to over-provide services (from a social point of view) as the negative consequences on the fellow practitioners' resource constraints are not taken into account. The implications of this for the provision of quality remain to be explored.

5.2 Salary

A salaried GP does not have to bear any financial cost in treating patients and only faces the non-monetary cost of effort. Again, the GP does not have a positive financial incentive to provide quality. A salaried GP has an incentive to distort the structure of services in a way that minimises effort. An efficient mix of services is, therefore, not guaranteed. A non-altruistic GP has an incentive to minimise effort by engaging in excessive referral and prescription. This leads to cost-ineffectiveness and to a reduction in the quality of care relative to what is possible with the resources expended.

As the GP's income is not linked to the demand for their services, (market) competition has no direct effect. An indirect effect of competition may arise if the organisation that employs the GP has to generate revenue in order to remain in business. In this case, an incentive may arise from the physician seeking to ensure the existence of their employer. Notwithstanding this, salaried GPs – just like other salaried employees – face very weak direct performance incentives.[38]

However, as I show in chapter 7 below, the payment of a salary or a budget may still be optimal if the GP is motivated intrinsically or by the quest for status. Indeed, it may in these cases be superior to relying on performance related payments.

5.3 Capitation

The incentive properties of capitation have been widely studied (e.g. Zweifel and Breyer 1997, section 7.3; Gravelle 1999; Chalkley and Malcomson 2000; McGuire 2000). Under capitation, GPs receive a fixed payment for each patient on their list. Thereafter, they have to bear the full monetary and effort cost of providing care for their patients. Under capitation, the GP has an incentive to contain cost by employing inputs efficiently. While cost reductions may also be achieved by curbing quality, the incentive to do so is held in check if demand, i.e. the number of patients on the list, increases with the quality of the service. In this case, the physician has an incentive to provide good quality services in order to increase the list size and, thereby, increase total income.

As already argued, the demand response mechanism guarantees the provision of appropriate levels of quality only to the extent that patient demand is sensitive to differences in quality. I have detailed a variety of reasons why this may not be the case. Most of them are applicable under capitation.

A list-based system is likely to be characterised by greater switching costs once a patient has registered with a particular GP. In some systems, patients are allocated to lists on administrative grounds and do not have a right to switch according to their own preferences. In this case, switching costs are obviously very high as technically a switch could only be achieved by a change of address. But even in systems, which allow patient choice of doctor, such as the UK, switching may involve an effort greater than any expected gain. Furthermore, the more stable relationship between patient and physician in a list-based

38 For an excellent introduction to the relationship between performance and payment systems, see Milgrom and Roberts (1992, chapter 7).

system allows for the sort of cumulative mutual experience the loss of which would constitutes a form of switching cost.

In a gate-keeping system, physicians are generalists who, until they refer to a specialist, administer a variety of services relating to all forms of ailments. However, they receive an undifferentiated capitation payment per patient, which does not reflect the differences in the marginal monetary and effort cost between the various modes of treatment that are available for a given condition.

Unless patient demand is sensitive to the overall quality of the service, physicians may be tempted to choose to administer those treatments that minimise their own cost rather than those that maximise the net benefit. Furthermore, they may focus on those aspects of treatments (convenience attributes) that are readily observable for patients and neglect those aspects that are difficult to observe (quality of clinical decision making). Finally, an undifferentiated capitation payment per patient does not reflect differences in treatment cost across patients. It therefore allows physicians to engage in discrimination against patients expected to incur high treatment costs, and under-provision of quality to patients who have a low quality elasticity of demand.

Example: Fundholding. In 1991, the then Conservative Government introduced an 'internal market' into the NHS, where purchasers contracted with providers (Le Grand et al. 1998). These reforms included a fundholding scheme, by which GP practices of sufficient size could opt for managing their own budget for purchasing certain secondary care services and pharmaceuticals. The idea was that for reasons of information, combined with a financial incentive, fundholders could be better purchasers of care than local Health Authorities, with regard to both efficiency and quality. As the fundholder's budget increased with the number of patients on their list, fundholding can be viewed as a form of capitation. Whether, in practice, it is different from a fixed budget then depends on whether the demand response mechanism is sufficiently strong to provide incentive effects. The following arguments relate to the effect fundholding on the quality of services. The evidence is mostly of a qualitative or even anecdotal nature (Goodwin 1998) and is inconclusive:

- *Referrals – Fundholders are likely to purchase secondary care more in line with their patients' need, implying higher quality. However, the incentive to economise on funds might lead to under-referral and, thus, lower quality. The evidence is inconclusive on whether or not fundholders were successful in incorporating quality standards into their contracts with providers of secondary care. The evidence suggests that providers have been more responsive to the demands of fundholders. Furthermore, access to secondary care may have been better.[39] However, some evidence also suggests that fundholders tended to refer less often than non-fundholders (see section 5.5 below).*
- *Prescriptions – A trade-off arises between a higher appropriateness of prescription against an incentive to under-prescribe for reasons of cost saving. The evidence suggests that cost saving incentives were only short lived.*
- *Primary care services – Fundholders may provide a greater range of services themselves, thereby enhancing convenience for patients due to local access. Administrative workloads of fundholders were higher, potentially at the expense of time spent in patient consultation. There is evidence supporting both of these points.*

The results by Dixon et al. (1997) reflect the ambiguous role of fundholding. They report that patients were more likely to leave practices with fundholding status; but were less likely to leave multi-funds (groups of practices holding a single budget). This might support the view that patients are concerned about (potential) rationing of services on the part of fundholders; but perceive the greater choice and variety of services within multi-funds as an element of quality.

39 On the role of primary care organisations as commissioners see also section 9.2 below.

5.4 Fee-for-service

One way to deal with the problems of selective quality provision and discrimination under capitation payments is to make fee-for-service (FFS) payments for those identifiable services or patients for which under-provision of quality is a particular problem. Indeed, recent years have witnessed a number of capitation based health services introducing FFS elements, including the UK and Denmark (Rochaix 1998). Such a strategy of selective FFS payments is only possible if the under-provision of quality is related to individual services or patients, which can be identified by the regulator. If this is not possible, the only way to enhance quality may lie in the introduction of a FFS system across the board.

In a FFS system, physicians receive a fee for each individual service they provide. Thus, revenue increases with the number of cases they see and the number of services provided per case. If services are produced at constant marginal cost and if physicians incur some fixed cost, physicians receive a positive income and engage in the provision of services only if the fee exceeds the marginal cost for at least some services. From this, there arises the well-known incentive for physicians to over-provide any service on which they receive a positive mark-up, i.e. a fee in excess of marginal cost (Zweifel and Breyer 1997, section 7.3; McGuire 2000). While over-servicing has been debated mostly in the context of its role in inflating health care expenditure, there are also implications for quality.

A purely income-maximising physician could prescribe a service up to a point at which the marginal effect on health becomes negative. Even if the benefit to health is positive, it may become small enough to be outweighed by the patient's time or travel cost or some other disutility from the treatment. Thus, over-servicing may well imply a direct reduction in quality and patient welfare. Secondly, unless the fee structure reflects the pattern of patients' marginal benefits, the services provided tend to be those that yield a greater margin rather than those bringing the greatest benefits to patients. Again, this implies a reduction in quality relative to the best outcome (Zweifel and Breyer 1997, section 7.3). Finally, as the practitioner's remuneration increases

with the volume of services provided but not necessarily with their quality, a FFS system may still fail to induce quality-improving effort. Indeed, over-provision of services is likely to lead to a deterioration of quality as the physician invests less time in the provision of each individual service.

Policy implications: While it is likely for services provided by a physician in person that an increasing marginal cost of effort places a natural bound on over-servicing, additional checks may have to be introduced by the regulator in the form of cost-sharing or global budgeting. Under cost-sharing, within a mixed payment system, the physician receives a practice allowance and/or capitation in addition to the FFS payment. If the allowance or capitation is set at a sufficient level, the fee for each service can be set below marginal cost. While the incentive for over-provision of services is, thus, eliminated, the demand response of patients to higher quality and/or a response by physicians to non-monetary incentives (e.g. intrinsic motivation, professional status) become the only guarantee of the provision of good quality. Mixed payment systems have been identified as a good compromise between the weaknesses of pure capitation and FFS payment (Rochaix 1998; McGuire 2000).

Some European health care systems (e.g. France and Germany) rely on global budgeting as a mechanism to curb over-provision in the presence of FFS. Here, a budget is fixed ex-ante and physicians can provide services at a fee for each service up to the limit at which the budget constraint becomes binding. From this point any additional service provided leads to a reduction in the fee – or an equivalent penalty. Suppose a budget has been fixed for some service and total expenditure on this service must not exceed the budget. The fee for this service is then fixed as long as the budget constraint is not binding. Once the budget constraint becomes binding, the fee is adjusted downwards for an increasing volume of service, so as to maintain budget balance. As the fee becomes a decreasing function of the service volume, the physicians' incentive to administer additional services is bounded, as they cannot enhance their income any further.

Little research has been carried out on the mechanisms and effects

of global budgets. As discussed in section 5.1, a violation of equity may arise when budgets have to be spent sequentially and efficiency may be compromised due to incentives to over-spend. The implications of this for the provision of quality remain to be explored.

So far, we have ignored the effects of the reactions of patient demand to quality under a FFS system. As far as quality can be observed by patients, their option to forego a physician's services altogether or abandon them in favour of a competitor, induces quality competition similar to that discussed in chapter 4. There are a number of aspects, however, which are particular to a FFS system.

First, quality competition under FFS places a bound on the over-provision of services if and only if over-provision leads to a reduction in quality from the patient's point of view. If, in contrast, patients receive a benefit from a greater volume of services, quality competition tends to exacerbate, from a social point of view, the over-provision of services. This problem has been extensively discussed in the context of the US hospital sector (e.g. Dranove and White 1994; Dranove and Satterthwaite 2000), but it is equally valid in the context of primary care. It follows that measures of cost-sharing or global budgeting are needed even more in the presence of competition between providers than they are in the case of a monopolistic physician. Furthermore, the level of (excessive) service expansion increases with the extent of the patients' insurance coverage.

Policy implication: One means of curbing over-provision of quality is to introduce patient co-payments, which expose the patient to at least some of the financial consequences of receiving more services. Co-payments, or deductibles, are extensively used by private insurers. The introduction of such measures within public sickness funds has been debated as an option in Germany.

Second, in contrast to capitation, which is linked to the presence of patient lists and often also of gate-keeping (Denmark, UK), FFS is usually linked to free patient choice and direct access to ambulatory specialist services (France, Germany). In as far as list based systems entail greater switching costs to the patient, one would expect quality

competition to be stronger in the non-list based FFS systems. Fleming (1992), for instance, attributes the extraordinarily high ratio of indirect to direct referrals[40] from GPs to specialists in Germany (91.3 as compared to 30.3 in the Netherlands or 1.6 in the UK) to the strong competition to which German GPs are exposed both with one another and with specialist practitioners.

It is well known from industrial organisation theory that firms have an incentive to evade competition by differentiating their products (e.g. Tirole 1988, chapter 6; Cabral 2000, chapter 12). Product differentiation implies that products are made poorer substitutes for each other from the consumer's point of view. Thus, under product differentiation each firm experiences a lower price elasticity of its own demand, as consumers are more reluctant to switch to a competitor's product in response to a price increase. One may expect similar service differentiation to occur in the market for physician services if this allows a reduction in the quality elasticity of demand and, thereby, a stifling of head to head quality competition.

The following data on the pre-tax income of physicians in private practice in Western Germany are indicative of the incentive to differentiate (European Observatory on Health Care Systems, 2000, Table 21). For 1996, average pre-tax physician income in general practice was DM 155,800, compared with average pre-tax income across specialist physicians of DM 200,600. Service differentiation appears to be taking place at a strong pace in competitive FFS systems, where many practitioners take on a specialisation or offer services relating to alternative medicine or physiotherapy. In the period 1990-1998, for instance, the number of office based specialists in Western Germany increased by 37%, whereas the number of GPs increased by only 14%. The share of GPs has thereby dropped to less than 40% of the practitioner population (European Observatory on Health Care Systems, 2000).

40 An indirect referral is defined as being initiated by a GP without a prior consultation with the patient. It, thus, implies more or less a direct access to the specialist. To the extent that patients value direct access, a high ratio between indirect and direct referrals reflects a strong degree of patient power.

Regarding the consequences of service differentiation, one can again draw a parallel with conventional industrial organisation theory. Even if consumers value some degree of product differentiation, it tends to become excessive from a societal perspective, as competition is stifled and investments in product differentiation are undertaken which merely transfer surplus from consumers to firms but do not generate additional value overall. For the market for physician services this would imply an excessive degree of service differentiation leading to a stifling of quality competition. Whether or not such a concern is valid remains to be tested empirically.

The presence of asymmetric information gives rise to a number of additional issues if patients have direct access to specialists, as is the case in many FFS systems. Specifically, the expert problem becomes prevalent (e.g. Wolinsky 1993; Emons 1997). This is because the treatment administered by a specialist is frequently of a credence nature, where patients cannot evaluate its quality. Moreover, they usually do not even know ex-post whether or not a particular treatment had been instrumental in curing their condition. Thus, both search and reputation fail to resolve the informational asymmetry.

A patient may instead try to overcome this problem by consulting several expert opinions. Wolinsky (1993) shows, albeit in a price-setting framework, that the experts may then specialise either in diagnosis only or in both diagnosis and treatment. Due to their lack of interest in inducing demand for treatment, 'diagnosis only' experts maintain their reputation for being honest (Dranove and Satterthwaite 2000). Thus, patients may visit a GP as a 'diagnosis only' expert and only upon their recommendation visit a specialist, who offers a second diagnosis and treatment. Wolinsky (1993) thereby offers an explanation for the presence of gate-keeping GPs even within a system, which in principle grants direct access to specialist services. While the informational asymmetry is thereby resolved, the cost of this is an excessive degree of diagnosis.

When patients are uncertain about the exact nature of their condition, the expert problem relates not only to the level of treatment but also to the choice of expert in the first place. A situation may arise in which a patient visits a practitioner who is offering treatment in

spite of a lack of qualification. In this regard, there may be under-referral between specialists. Obviously, a quality problem arises when patients are being treated by a poorly qualified specialist. Lu et al. (2000) show how this mismatch problem can be overcome by the inclusion of a quality related performance component into the providers' payment. In such a case, providers can raise income by referring on the patients they are unqualified to treat and so are unlikely to bring to a good outcome. Lu and colleagues substantiate this finding empirically by analysing the effects of the US state of Maine's performance related payment system for substance abuse treatment.

Finally, in systems, such as the German, with a mix of private and public health insurers who strike different payment contracts with physicians, an issue may arise about discrimination according to insurance status. Specifically, practitioners have an incentive to distort quality upwards (downwards) for those patients, for whom they can expect more (less) generous fees. In Germany, this generally implies that privately insured patients may receive preferential treatment.[41]

5.5 Empirical evidence on the effects of payment systems

A substantial body of empirical work exists, which studies the impact of payment systems on physician incentives (for an overview, see Scott 2000, section 6; Gosden et al. 2001). Krasnik et al. (1990) study the effects of a change from pure capitation to a mixed payment system including FFS elements, as the mode of remuneration for GPs in the city of Copenhagen. Using three data points, one before the change and two after, and a control group of GPs outside the city (Copenhagen county), where no change in the payment system had occurred, they employed maximum likelihood methods to estimate changes in GPs' activity. Their results confirm that the change from capitation to FFS raises the intensity of services provided by GPs

41 Dranove and White (1994) report a number of studies confirming discrimination according to insurance type in the US hospital market.

themselves, both diagnostic and curative, but decreases both referrals and prescriptions.[42]

Iversen and LurÂs (2000) studied the effect on referral rates of an experimental alteration in the remuneration of Norwegian GPs, where a practice allowance-cum-FFS payment was replaced by a capitation-cum-FFS payment. In this case, the theoretical prediction is not immediately clear, as both capitation and the practice allowance tend to provide incentives for referrals. Yet, for the following reason, capitation should be expected to give rise to greater referral rates than a practice allowance. GPs can use referrals to reduce the intensity of their service provision and, thereby, acquire the time to provide services to additional patients. In so doing, they enhance their income by capturing additional capitation payments. This incentive does not exist with a flat practice allowance. Iversen and LurÂs empirically confirm their hypothesis of greater referrals under the capitation-cum-FFS payment by analysing referral data for 33 GPs from two periods: before and after the change in the remuneration system. They estimate that the change in payment method gave rise to a 42% increase in the referral rate.

Giuffrida and Gravelle (2001) studied the demand for, and supply of, GP night visits in the UK before and after the introduction of a fee differential between GPs and deputies. The model predicts that GPs would respond to the fee differential by substituting their own for deputies' visits and that this response is magnified by demand management practice. Both predictions were confirmed empirically based on the analysis of 1984/85-1994/95 panel data for UK primary care.

Finally, recent empirical research has used panel data from the UK Health Authorities' Contract Minimum Dataset to (re-)assess the incentives under fundholding. Gravelle et al. (2002b) find for the case of cataract admissions to hospital within a large Health Authority that

42 Gosden et al. (2001) report the results of a systematic review of studies on the effects of remuneration on physician behaviour that were undertaken before 1997. They emphasise that, while most of these studies indicated behavioural responses to the remuneration method that were supportive of economic theorising, the methodological quality of all of the studies was so low as to put in doubt their validity.

fundholders yield lower admission rates than non-fundholders and that they respond differently to changes in waiting times and patient characteristics. Dusheiko et al. (2002) confirm the notion that fundholders tend to refer patients to specialists less often than do non-fundholders. They demonstrate empirically that the abolition of fundholding in the UK in 1999 reduced the difference between ex-fundholders and ex-non-fundholders in rates of hospital admissions for elective surgery but not in emergency admissions. In contrast, Juarez et al. (2002), who consider data from a different Health Authority, report that fundholders had higher admission rates than non-fundholders overall.

While all of these studies provide clear evidence for the relevance of the payment system for GP behaviour, none of them has explicitly linked it to quality indicators. Thus, it remains unresolved what the observed responses of physician behaviour to the payment system may have implied for quality, however measured.[43]

43 A number of empirical studies assess the effect of changes in US hospital reimbursement on a variety of quality indicators. The evidence is highly mixed. For a survey see Dranove and White (1994) and Dranove and Satterthwaite (2000).

Summary

- Fixed budgets generally imply an efficient mix of services but may entail an under-provision of services and quality if the physician is mainly motivated by financial concerns.
- If physicians are mainly motivated by financial concerns, a flat salary may lead to under-provision of effort and distortion in the mix of services towards referrals and prescriptions.
- Capitation gives rise to appropriate quality incentives if and only if the demand of all patients is responsive to all of the relevant quality dimensions. Otherwise, and if physicians are mainly motivated by income, capitation might lead to the under-provision of quality dimensions that are unobservable, or to discrimination between patients.
- Fee-for-service (FFS) tends to lead to over-provision of services. This has ambiguous implications for quality. Quality may be too high, i.e. at levels that are no longer cost-effective. Conversely, quality may also be too low, due to a distorted mix of services. For example, FFS payment may induce physicians to treat patients themselves even when a referral would be more appropriate.
- If patients have direct access to specialists this tends to provide an incentive for physicians to specialise in order to differentiate their services. This is supported by evidence from Germany.
- Empirical evidence supports some of the theoretical predictions about the incentives provided by different payment systems. However, little is yet known empirically about the implications for the quality of care.
- Between 1991 and 1999, some GPs in the UK were fundholders and used capitated budgets for the purchase of secondary care and medicines. Evidence on the effects of fundholding is mixed. Some studies, but not all of them, imply that fundholders tended to refer and prescribe less. However, once a referral was initiated, speed of access to secondary care appears to have been better for patients of fundholding GPs.

6 REGULATION

It is not easy to separate the role of regulation from other determinants of quality provision. The imperfection or absence of markets means that health care systems are replete with regulatory interventions, including the design of payment systems, the setting of quality standards, the introduction of audit systems, or the publication of performance indicators. The previous chapter on 'market' incentives has already addressed two sets of regulatory measures. We have discussed the role of those payment systems (practice fund, salary, capitation, FFS) which are not directly linked to quality measures. Moreover, we have discussed regulatory means, such as practice licensing or performance indicators, by which the regulator can mitigate the problem of asymmetric information in health care markets.

This chapter is devoted to regulatory incentives that are more directly linked to the provision of quality. In particular, we address the scope for the regulator to influence quality provision by means of quality related performance pay and by the introduction of quality standards. Implied by this is the introduction of some form of monitoring system. The discussion in the current chapter is framed in the context of 'hard' regulation – financial incentives or legally binding standards – being imposed upon the GP by the payer. I still omit from the discussion two important issues of regulation: the non-income related effects of regulation on physicians' motivation, which we address in the next chapter; and the important role of self-regulation and clinical governance, which I address in chapter 8.

In regard to the issue of external regulation versus self-regulation, it should be noted that in the present chapter I follow the literature on regulation in assuming that it is the regulator who determines the payment schedule. This assumption can, of course, be debated in the light of the practice of joint bargaining over fee-schedules between health care payers and physician associations. The issue of payer-provider negotiations is of importance and has, at least in the European context, received little attention from either theoretical or empirical researchers.[44] Nonetheless, I will not pursue this topic any further, if only for the reason that the implications of corporate bargaining for the provision of quality remain to be explored.

6.1 Quality-related performance pay[45]

I showed in chapter 4 that fee payments give rise to proper quality incentives if and only if patient demand is sufficiently reactive to all relevant dimensions of quality. If this condition is not fulfilled – a likely instance – or if physicians receive a fixed budget or salary, then the regulator has to provide more direct quality incentives. This may be achieved by linking the GP's personal compensation, i.e. salary or capitation, to some measure of quality. In so doing, the regulator establishes a link between quality and the physician's income. Such forms of payment have recently found their way into American health care systems (Lu et al. 2000).

Example: Within the reformed English NHS there is an expectation that Primary Care Trusts will devise incentive schemes for their constituent practices with a view to meeting national targets (Department of Health 1999). Incentives include the following:

- *non-consolidated cash bonuses for individuals and teams;*
- *paying for additional equipment or a one-off investment [...] as a reward for a high-performing team;*
- *non-recurrent expenditure on providing improved support services for key staff;*
- *education, training and personal development;*

44 Zweifel and Eichenberger (1992) have studied the empirical relevance of cartelisation within health care systems (see chapter 8). Gravelle (1999) studies a physician cartel within his spatial model of GP competition. However, due to the specific nature of the cost function little can be said about the implications of cartelisation for quality. Demange and Geoffard (2002) study within a theoretical model the difficulties a policy-maker faces when trying to reform the reimbursement system subject to having to win over political support from physicians. Again, quality issues are not addressed.

45 The following section draws heavily on the general principles of performance pay laid out in Milgrom and Roberts (1992, chapter 7). See also Chalkley and Malcomson (2000) on performance pay in health care contracts in general and Goddard et al. (2000) on the economics of the NHS performance framework and some of its pathologies.

- *setting up a fund for providing additional training or development for key staff;*
- *improving the physical environment for key staff as a reward for delivering high performance (Department of Health 2002b).*

While it is pointed out that practices will have to 'earn' their entitlements to rewards, it remains unclear whether some of these 'rewards' should not also be targeted at poor performers with a view to enhancing the skill base (education, training and personal development) or improving the working environment. This would obviously undermine the incentive basis.

One of the instruments to provide incentives to practitioners is the introduction of quality-related components into Personal Medical Service contracts (Department of Health 2000a, 2003). It is also envisaged that elements of performance pay will be a major part of a new the General Medical Services contract at national level.

For the following, suppose that the GP receives a salary or budget, so that a priori there are no financial incentives to provide effort. A performance payment should ideally be a direct function of the GP's effort. This assumes that the payer (the principal) has a way not only of monitoring the physician's (the agent's) effort but also of objectively verifying it. The realistic assumption of most principal-agent models is, however, that this is impossible (e.g. Milgrom and Roberts 1992, chapter 7).

The principal must therefore link the payment to an observable and verifiable signal, which is correlated with the GP's effort. Such signals may relate to the process of delivery or to health outcomes. For example, a payment may be linked to the level of an important service provided by the GP, such as payment for vaccination or cancer screening.[46] The payment may also be linked to some measure of health outcome (e.g. Zweifel and Breyer 1997, section 8.2). The incentive payment then induces greater effort if the agent can thereby improve the expected outcome and increase the payment received.

46 Note that both FFS and capitation payments can be interpreted as performance pay, with the observed signals being service volume and list size, respectively.

The problem is that the aforementioned signals are only imperfectly correlated with true effort and are subject to random influences or to manipulation. Let us now address some of the consequences.

6.1.1 Multi-tasking and the equal compensation principle

As we know from the discussion of FFS systems, a performance payment related to the volume of a service may induce the physician to over-produce this particular service, leading to sub-optimal provision of quality. Thus, even if such a payment induces additional effort in the delivery of the service, it may discourage the GP from providing effort in other dimensions of care. A similar problem arises if the payment is based on an outcome measure embracing only a subset of the relevant quality dimensions. Again, there is a danger that the GP focuses effort only on the dimensions of quality that are being monitored and rewarded, while neglecting others. If the agent has to perform multiple tasks, incentive payment can induce an optimal delivery of the overall service only if effort in all tasks is being rewarded. Milgrom and Roberts (1992, chapter 7) call this the 'equal compensation principle'. Due to the distortion in the provision of effort, a partial reward of individual tasks can easily lead to an overall outcome worse than it would be in the absence of performance pay.

Policy implication: When devising quality-related performance-pay the regulator has to be confident about reimbursing all of the relevant dimensions of quality. Severe constraints may arise due to the impossibility of measuring and/or reimbursing some dimensions such as clinical effort or empathy in the physician-patient interaction.[47]

6.1.2 Incentives in teams

Health care for a single patient is usually the outcome of a joint production process involving a number of different physicians, nurses and other health care professionals. This is most obvious for the case of referrals, where both primary and secondary care physicians are

47 In this regard, recall the problems associated with performance indicators (section 4.4).

involved in the production of care. In such a case an issue arises about whether incentives should be team-based and relate to the outcome or whether they should be targeted at the individual contributions.[48] The crux is that with team-based compensation individuals cannot acquire the full returns to their individual effort and, therefore, tend to under-provide it. The problem with rewarding individual contributions is that the provision of the required co-operation cannot usually be contracted for, implying that incentives are missing for an important element of performance.

Policy implication: The well-documented possibility of failures of communication between GPs and secondary-care specialists (Wilkin and Dornan 1990; Jankowski 2001) highlights the importance of team-related incentives; in this case at the primary-secondary care interface (see also section 9.2).

6.1.3 Risk and the incentive-intensity principle

We have already argued that outcome and process measures are likely to be subject to uncertainty. Health outcomes are determined by a variety of factors, which lie beyond the immediate influence of the primary care physician. These include environmental factors, effort on the part of secondary care providers, the patient's compliance with the proposed treatment, as well as pure chance.[49] Thus, even if physicians put in their best effort, a bad outcome may arise; likewise, poor effort may yield a good outcome just by chance (Zweifel and Breyer 1997, section 8.2).

The ongoing debate about the validity of outcome research, i.e. the rigorous determination of which treatments are effective, (e.g. Tannenbaum 1993; Berrow et al. 1997) and the difficulty of performance measurement (e.g. Blumenthal and Epstein 1996; Giuffrida et al. 2000; Goddard et al. 2000) suggests the presence of

48 See Ratto et al. (2001) for an application of the theory of team incentives to the NHS.

49 Giuffrida and Gravelle (1998) analyse how the GP (as principal) may induce the patient (as agent) to comply with certain measures of treatment. Thus, patient effort may be partly within the GP's control.

considerable measurement error as a second source of uncertainty. Random variations in measured outcomes expose a GP who is subject to performance pay to an income risk. Risk-averse GPs can only be induced to take up the profession if they are compensated for this risk by an outcome-independent premium that increases with the degree of risk and the degree of the GP's risk-aversion. Otherwise, the risk of appearing to under-perform may induce the GP to engage in defensive medicine in the form of over-diagnosing, over-treating or over-referring patients to specialists.[50] In order to contain the risk premium as well as risk-related distortions in the GP's treatment choices, the regulator has to reduce the incentive component of the payment and, thereby, reduce the risk to which the physician is exposed.

Policy implication: The incentive intensity of the performance payment and, thus, the stimulus to provide quality should decrease both with the level of risk associated with a particular outcome measure and with the GP's level of risk aversion. The presence of substantial risk implies that, in spite of its inferior incentive properties, performance payment which is based on inputs and/or process measures may prove to be superior to outcome based payment.

6.1.4 Relative performance evaluation and the informativeness principle

Frequently, the principal has available more than one signal of the agent's effort. For instance, the principal may observe elements both of the treatment process (e.g. the number of flu vaccinations) and of the outcome (e.g. the number of flu cases). The principal should then relate payment to those signals in a way that minimises the agent's risk. Specifically, this implies that the principal should base payment on those measures that are subject to lower variance.

50 Similar problems arise in the presence of malpractice litigation. This appears to be more relevant in the context of secondary care. See Danzon (2000) for an extensive overview of the subject.

Policy implication: If clinical outcome measures are highly variable (e.g. the prevalence of flu), then performance pay should rather be related to process measures (e.g. the number of vaccinations) even if this induces the physician to over-treat.

Example: Drawing on evidence on learning-by-doing in the provision of health care, the German Advisory Council has proposed linking specialists' pay to their 'routine' as measured by a minimum number of procedures performed. Routine might be a better measure for a physician's ability than measures of outcomes. However, they also caution that appropriate safeguards have to be provided against the physicians' incentives to over-provide procedures in order to claim the bonus (Sachverst‰ndigenrat 2001, section 3.2).

Furthermore, the principal may be able to reduce risk by using signals that are not directly related to the agent's effort, but which are correlated with the deviation of the realised outcome from the expected outcome. For illustration, consider the following example. Suppose GPs are remunerated according to their success in reducing the number of emergency hospital admissions due to acute asthma conditions. Thus, a GP's remuneration falls as this number increases. Clearly, such a payment exposes a GP to risk since despite their best efforts, adverse environmental conditions, such as periods of intense smog or epidemics of acute respiratory illness, may give rise to an unforeseen boost in emergency emissions. The regulator could reduce the GP's risk by either of the following two means.

First, the principal could make the performance pay contingent additionally on a direct signal of the environmental conditions, e.g. air quality, weather conditions or the general prevalence of acute respiratory illness. Here, possible reductions in a GP's pay due to increased emergency admissions are compensated if the direct signal indicates the presence of adverse environmental conditions such as a flu epidemic or a period of smog.

Second, the principal could make the GP's remuneration contingent not on their absolute success in reducing emergency admissions but rather on their success relative to other GPs working

under similar conditions. Thus, the payment becomes a function of the number of emergency admissions from the GP relative to the average. If adverse conditions affect all GPs, the own admissions and average admissions will rise in tandem and thus leave the GP's pay unaffected. This form of relative performance pay, sometimes dubbed 'yardstick competition', filters out the risk that is common to all GPs and, thereby, contributes to a reduction in the GP's overall payment risk.

Policy implication: Wherever a group of GPs is exposed to the same risk, relative performance pay is superior to absolute performance pay.

The general principle evolving from these arguments is that all signals should be used which reduce the variability of the agent's compensation. This is known as the informativeness principle.

6.1.5 Monitoring intensity principle

If monitoring were costless, the principal should be able to induce appropriate effort on the part of the GP simply by observing each and every move and penalising any defection from best practice. In reality, however, an agency problem arises from the presence of monitoring and enforcement costs. Indeed, the cost of monitoring and enforcing best effort is likely to be prohibitively high. Monitoring of process or outcome signals usually involves a substantial cost too. The principal is willing to incur monitoring costs only if a significant increase in performance can be expected. It is then easy to see that it only pays the principal to engage in monitoring if the incentive-intensity is high. This is because by reducing the measurement error, the principal reduces the GP's risk and, thus, the risk premium. Conversely, if the incentive component in the GP's payment is weak the risk premium is also negligible. In this case, there is no need for the principal to engage in costly monitoring.

Policy implication: Systems of performance pay should be accompanied by appropriate means of monitoring the relevant parameters. This implies, for example, that clinical audit has to be designed carefully in a manner that optimises informativeness.

6.1.6 Determining the benchmark: adverse selection and the ratchet effect

So far, I have said nothing about how the regulator can determine the benchmark against which to value the GP's performance. Such a benchmark is important in determining the overall level of fixed and performance based remuneration. If payment is linked to an outcome measure, the benchmark may be set at the outcome level that can be expected if the GP provides an optimal quantity of effort. However, the regulator is unlikely to know this level. While clinical trials can produce evidence as to what outcomes can be expected, contingent on case mix and mode of treatment, the expected outcome for individual GPs still depends on the case mix they face as well as on their personal skills. With these characteristics not being readily known to the regulator, performance pay may be wrongly tuned. Thus, a GP's pay may either be excessive or it may fall short of what is necessary to guarantee ongoing participation in the profession.

The regulator is confronted with a problem of making a contractual offer while having incomplete information about the GP's type. This leads to a problem of adverse selection akin to the one discussed in section 4.3. The literature suggests that this may be overcome by means of complicated contracts that induce self-selection, but usually at the cost of a loss of efficiency (reviewed in Chalkley and Malcomson 2000).

A second possibility lies in the use of past performance as a benchmark. Here, the so-called ratchet effect may cause a problem. Suppose that the regulator observes good performance in one period and consequently 'ratchets up' the benchmark for the following period. Then, good performance in the first period is effectively punished. Anticipating this, the GP has an incentive to under-perform to begin with in order to benefit from a lower future performance target. As a result of such gaming, under-performance becomes endemic to the system.

Policy implication: The imposition of performance standards as benchmarks is important in order to fine-tune performance pay. However, benchmarking is prone to gaming under imperfect information.

6.1.7 Rent extraction and specific investment

A related problem with the design of compensation lies with the trade-off between leaving the GP as little rent as possible, on the one hand, and inducing appropriate investment in human and technological capital, on the other. Physicians invest only if they expect an adequate return. One return to an investment in diagnostic skills or equipment may lie in the GP's ability subsequently to achieve any given performance with less effort. With performance pay, the doctor receives a rent in the sense of attaining a level of payment above the appropriate reimbursement of 'true' effort. Generally, the payer has an interest in extracting these rents, e.g. by ratcheting up the performance target. The problem is that by extracting the rent the payer strips the GP of the returns to investment and hence removes the incentive to invest.

Many medical skills are highly specific and, at least in public health services, there are few outside opportunities to work as a physician. Thus, as soon as GPs have undertaken ('sunk') an investment in medical skills, the payer might wish to reduce pay to the level at which GPs' are indifferent between remaining in the service and leaving for some outside profession from which, for lack of proper training they can only expect modest returns. GPs would, in that case, effectively have been 'held-up' by the payer.[51]

Rent extraction by the payer implies that GPs are stripped of the returns to their investment. In as far as they rationally anticipate being 'held up' in this way, they will under-invest in medical skills or not enter the profession in the first place.

51 The form of ex-post contractual opportunism described by the 'hold-up problem' has been widely discussed in the fields of industrial organisation, regulatory economics and organisational economics. See, for example, Milgrom and Roberts (1992, chapter 5) or Chalkley and Malcomson (2000).

Policy implication: If prospective physicians expect to be 'held up' by the regulator, under-qualification becomes endemic and a sub-optimally low level of quality is provided. To avoid this problem requires the payer of physician services to acquire a credible reputation for maintaining fair remuneration.

The problem with 'hold-up' is that, even if the payer announces a satisfactory level of pay before the future GP invests in medical capital, it is always optimal for the payer to break this promise later and extract the rent, once the physicians have made their investment in careers as GPs. This undermines the payer's credibility when announcing high pay schedules in order to attract qualified candidates and stimulate investment in skills.[52] The frequency of health care reforms and the urgency of cost reductions in most of today's health care systems cast some doubt as to whether such a reputation can be established, however.

If quality incentives arise neither from the market nor intrinsically, performance pay may be the only way to elicit appropriate levels of GP effort. The discussion in the preceding paragraphs should have shown, however, the difficulties in designing appropriate incentive schemes, where multi-tasking, risk, imperfect monitoring and the principal's inability to commit to 'fair' payment undermine the scope of regulatory quality incentives and physicians' incentives to invest in medical skills.

6.2 Clinical guidance and variations in practice

Recent health care reforms, in particular in the UK's NHS, have focused on setting general standards of care as a means of reducing variation in clinical practice. Clinical guidelines developed by the National Institute for Clinical Excellence (NICE) constitute the basis for national standards for key conditions, which together with access standards are laid down within National Service Frameworks (NSFs)

52 One way to extract rent may be the ex-post stimulation of additional entry into the GP market, where rents are extracted by way of greater competition.

(Department of Health 2000a, 2002a). The guidelines contained within the NSFs are detailed and apply to both secondary care and primary care (see for example the NSF for Coronary Heart Disease, Department of Health 2000b).

In particular, there is concern about wide variation in referral rates (Wilkin and Smith 1987).[53] Most research has found it difficult to explain a substantial amount of residual variation in medical practice other than by variations in the individual practitioner's style (Wilkin 1992; Davis et al. 2000; Phelps 2000). Phelps (2000) argues that practice styles develop as reactions to medical uncertainty and depend on the medical schooling received by the practitioner in question.

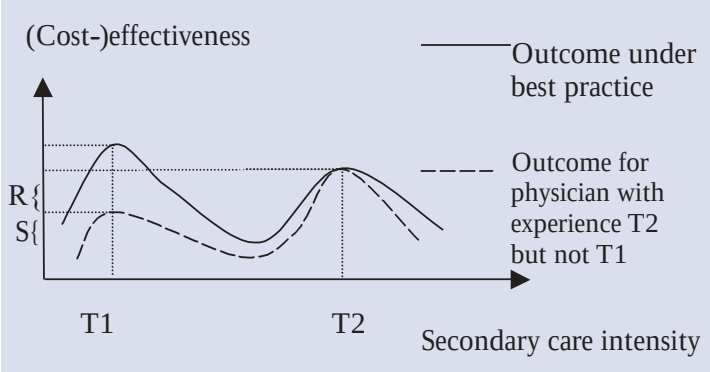
There is widespread agreement that unexplained variation is to some extent a symptom of inefficient, ineffective or even harmful practice, as well as a symptom of the presence of inequity in health care. It is equally accepted that there are benefits to the dissemination of information about best practice, e.g. in the form of guidelines or research and dissemination networks (Phelps 2000; Thomas et al. 2001). It is less clear, however, whether the imposition of 'hard' targets or standards, as advocated by some health care administrators, can improve upon the situation. To begin with, it has to be established what constitutes best practice. With regard to referrals, for example, the question of best practice remains unresolved for many conditions (Roland 1992).

For an illustration of the problem, consider the following example, which is depicted in Figure 6.1. Suppose that treatments can be characterised by the volume of secondary care they entail. This is shown as secondary care intensity on the horizontal axis of Figure 6.1. For example, treatment T1 relies on primary care inputs whereas treatment T2 is secondary-care-intensive. The vertical axis describes the outcome in terms of (cost-) effectiveness for different treatments depending on their secondary care intensity. Under best practice (the solid curve) the treatments T1 and T2 constitute 'local' optima in that they dominate all other treatments in terms of their outcomes. Assume

53 Wilkin and Smith (1987) report a mean of 6.1 referrals per 100 consultations, with a range from 1.0-24.0, for a sample of 201 UK GPs.

now that each of these treatments is being applied by a number of GPs, who have varying degrees of proficiency in them. The following observations can then be made.

Figure 6.1: Variations in outcome



To begin with, it is clearly inefficient to reduce variation by way of forcing treatment practice towards the mean. Any 'average' treatment, corresponding to any point on the solid curve between T1 and T2, is clearly less desirable in terms of outcome than either treatment T1 or T2.[54] This implies a strong caveat for standards that are developed on a consensual basis and so are likely to be biased towards the mean.

54 Wilkin (1992) argues that establishing an intermediate referral standard may give rise to cost increases if different referral patterns arise from differences in case-mix between GPs.

Now suppose that outcomes research identifies the primary care based treatment T1 to be superior in terms of outcome as illustrated by the global peak of the best practice curve in Figure 6.1. This clearly justifies a focus on treatment T1 in the education of new physicians. It is less clear, however, whether it would be beneficial to introduce a standard imposing on all physicians the use of treatment T1.

Given their historical medical education, some practitioners may have developed strong experience in administering treatment T2. Forcing these practitioners into administering treatment T1 may, at least temporarily, give rise to a reduction in the quality of their care well below the level they would achieve with treatment T2. This is illustrated by the broken curve in Figure 6.1, which represents the outcome attained by a physician who is trained in T2. It is reasonable to presume that the physician's performance deviates from the best-practice outcome as follows. Due to lack of experience, the physician under-performs when administering any other treatment than T2 (the broken curve lies below the best practice curve).

Moreover, performance is likely to deteriorate the more radically different treatments are from T2 (the gap between the solid and broken line increases the greater is the distance from T2). Hence, it cannot be ruled out that a physician trained in T2 achieves an outcome for the 'best' treatment T1 below that which they achieve with treatment T2.

Such a situation is particularly likely in the case of complex treatments involving a range of complementary inputs and skills.[55] In this case, a shift from treatment T2 to treatment T1 does not merely require substitution of primary for secondary-care services but rather a complex adjustment across the whole range of inputs.[56] This may imply that the quality reduction would last for an extended period of time. If the difference in effectiveness of the treatments T1 and T2 is not too large (distance R), the cost of retraining older GPs with

55 Two inputs A and B are complements if output can be raised for an increase in input A only if input B is increased as well.

56 The argument here is very much akin to the problem of switching production processes in manufacturing industries in the presence of strong complementarities (Milgrom and Roberts 1990).

experience in T2 and the short-term loss in patient welfare (distance S, S>R) may outweigh any longer-term gains. Furthermore, while variation in referral patterns is reduced by focusing on T1 as the recommended treatment, variation with regard to outcomes may increase (S>R).[57] Arguably, it could then be better to allow practitioners trained in treatment T2 to carry on referring at a higher rate and phase out treatment T2 by focusing the education of new GPs on treatment T1, rather than to require older GPs to retrain.

Policy implication: While the problem of variation in practice is likely to be substantial, regulators should be wary of the potentially harmful side-effects of 'hard' regulation, such as clinical guidelines. In particular, if practitioners differ in their skill base and expertise, the enactment of inflexible guidelines on best practice may lead to worse performance by some practitioners relative to the status quo. 'Soft' policies promoting the dissemination of information and the encouragement of continuous professional education may sometimes prove to be superior.

57 See Dawson et al. (2001) for a similar discussion of the impacts of performance standards on variation in quality of, and access to, secondary care.

Summary

- Recently, there have been moves in some health care systems towards introducing more direct quality incentives into physician remuneration. The advantage of performance pay lies in its provision of direct quality incentives even if patient demand is unresponsive to quality.
- The design of performance schemes involves a variety of problems that relate to:
  - the requirement that all important dimensions of performance need to be reimbursed;
  - the provision of team incentives;
  - the containment of performance risk to which the physician is exposed;
  - the appropriate degree of monitoring;
  - the determination of performance benchmarks; and
  - the regulator's credibility when committing not to extract the physicians' rent by ratcheting up standards.
- Concerns about variation in practice, e.g. with regard to referral behaviour, lead to the introduction, and sometimes the imposition, of practice guidelines. If physicians differ in their abilities and expertise, imposing the same practice on all may compromise the provision of quality by some physicians. In that case, dissemination of information and encouragement of continuing education may be a superior approach.

7 INTRINSIC MOTIVATION AND SOCIAL INTERACTION

The discussion of income based quality incentives has shown that both market and regulatory incentives are likely to be weakened or distorted by various imperfections, many relating to imperfect information. Indeed, it is doubtful whether some poorly observable dimensions of quality would be provided at all if physicians were merely interested in income. In this regard, it is fortunate that professional behaviour is not exclusively driven by financial incentives but also by status seeking, intrinsic motivation and altruism (Pauly 1980; Dionne and Contandriopoulos 1985; Lerner and Claxton 1994; Encinosa et al. 1997; Scott 2001).[58]

This chapter deals with non-monetary quality incentives. It also addresses the interaction between monetary and non-monetary incentives, which may pose a further caveat for the market or regulation as instruments for enhancing quality. A number of policy recommendations follow.

7.1 Altruism and payment incentives

Altruistic providers care directly about patient welfare, which, in the absence of patient co-payments, is unambiguously an increasing function of the quality of care that patients receive (Ellis and McGuire 1986, 1993; Lerner and Claxton 1994; Chalkley and Malcomson 1998b, 2000). In this regard, GPs may be acting in the interest of their patients but not necessarily in the interest of society. This is because as long as insurance or tax funding insulates patients from the cost of health care, they face a moral hazard to over-consume it, both in terms of quantity and quality.

Ellis and McGuire (1993) show that in this case, a mixed reimbursement system in which the regulator chooses the appropriate degree of cost sharing can achieve a socially optimal provision of care. The degree to which cost is shifted to the provider increases with the

58 The bulk of the literature on non-financial incentives for physicians is based on anecdotal evidence. Scott (2001) is an exception in using an experiment to show that non-monetary factors such as working hours, special interests, the use of guidelines and list size influence GPs' (location) choices.

degree of altruism. Whereas, in the absence of other incentives, selfish providers can only be induced to deliver quality by having their costs fully reimbursed, altruistic providers may have to be exposed to some or even the full cost of care in order to deter over-provision.

Chalkley and Malcomson (1998b, 2000) modify this argument for a setting in which the provider can additionally engage in cost-reducing effort. In this case, a fully prospective payment, i.e. capitation payments or a budget, induces an optimal cost-reducing effort but also an under-supply of quality. Starting from this point, the introduction of some amount of cost sharing between the payer and an (imperfectly) altruistic provider always induces a quality improvement towards the socially desirable level. However, this occurs at the expense of effort into cost-reduction.

It has been argued that physicians' altruism may embrace social welfare in addition to individual patients' wellbeing (Blomqvist 1991). In principle, this would allow for an alignment of practitioners' and patients' interests with those of society, and resolve the aforementioned moral hazard problem. However, this form of altruism is subject to a serious caveat. Generally, the cost saving effort of an individual GP is so small relative to societal health care expenditure as to be negligible. Anticipating this, a rational GP will then not cut back on treatment quality, as this gives rise to a tangible reduction in the GP's, as well as the patients', benefit without any significant offsetting gain to society. This is an example of the free-riding problem in the use of common property resources. As all GPs rationally engage in the over-provision of services, cost-savings remain intangible.

7.2 Intrinsic motivation and external incentives

It is generally accepted that professionals, such as GPs, are to some extent motivated by the satisfaction they receive simply from doing their job and doing it well. This phenomenon has been labelled 'intrinsic motivation' by psychologists (Deci and Ryan 1985). Frey (1992, 1997) has analysed the relationship between intrinsic motivation and external incentives arising from the price system,

regulation or supervision. Intrinsic motivation is likely to depend on the GP's autonomy and self-esteem. Frey (1992) proposes that both direct regulatory intervention and price incentives tend to reduce the marginal intrinsic benefit and, thus, at the margin, tend to crowd out motivation.[59] This is because both incentives and regulation tend to destroy the GPs' self-evaluation of doing 'something decent' over and above what is externally rewarded or enforced.

The loss of intrinsic motivation is even more pronounced under regulation as, in this case, the GP additionally suffers a reduction in self-determinedness.[60] Surprisingly perhaps, even the offering of prizes and rewards can sometimes destroy intrinsic motivation. This is the case if the GP gets the feeling that the reward is not being given in recognition of competence and professional ethos but rather in reaction to the achievement of an expected target. The prize would thus reduce the GP's self-esteem. In contrast to this, a form of 'soft' regulation or a 'moral appeal' may enhance intrinsic motivation by raising a compliant GP's self-esteem. Likewise, the publication of good practice – standing alone from any prior targets – tends to enhance intrinsic motivation. These arguments notwithstanding, the introduction of external incentives may be justified if the direct incentives are stronger than the crowding-out of intrinsic motivation.[61]

59 Kuhn (2001) studies the optimal allocation of a budget and professional autonomy to intrinsically motivated agents in the presence of asymmetric information. The problem is that the allocation should be designed such that the only agents who self-select 'high autonomy' are those who are efficient in using their own budget in producing a high value (quality) output. As it turns out, contracts that induce such self-selection may involve substantial distortions in the budgets and may become entirely unfeasible if the agents' preference for autonomy is sufficiently strong. The model may be applied to the new contractual arrangements for GPs within the UK NHS, which allow them to choose between contractor status, with a greater degree of autonomy and a self-administered budget, and salaried employed status with less autonomy.

60 Although empirical evidence is still scant, the issue of 'unhappy doctors' whose work morale suffers under an excessive load of regulation has recently been recognised as an issue for academic and policy attention (Edwards et al. 2002; Ham and Alberti 2002).

61 Barkema (1995) studies empirically the impact of performance incentives in Dutch business firms. He shows that the effect of external intervention on work performance is significantly positive (negative) in the case of impersonal (personal) control. Since intrinsic motivation tends to be more sensitive in personal relationships, the evidence lends some support to the crowding out hypothesis.

7.3 Social interaction and performance payments

Intrinsic motivation is effective even if agents operate in isolation, implying a certain degree of self-reference and self-reflection. However, a further source of motivation arises from social interaction. Here physicians do not compare their own performance against an ideal, but receive social feedback – social status – on the basis of their performance as compared to a social norm. If status is determined with reference to a fixed norm of performance or income, a status related incentive arises, which is not subject to competition. When choosing effort, GPs take into account the effect this has on their professional standing. The acquisition of status is then similar to the acquisition of a reputation, with the difference being that the GP receives a direct benefit from status.

What exactly determines status depends on societal attitudes. Status may be related to income and, thus, provide only an indirect quality incentive. But it may also arise from a reputation for high quality provision and, thereby, provide a direct quality incentive.

In many instances, social status is determined within a reference group, i.e. within a group of GPs. Thus, status becomes relative and is, therefore, contestable. Encinosa et al. (1997) study a setting in which physicians compete for status with regard to income and effort. The model could be easily expended to include status arising from clinical performance. Increased competition for income-related status or for professional status elicits additional effort and, thereby, leads to an enhancement of quality. However, if clinical performance is subject to random influences, the GP is exposed to a status risk akin to the income risk discussed in section 6.1. Encinosa et al. (1997) show that the additional status risk weakens the intensity of formal incentive schemes even below the level resulting from the presence of income risk.

Taking a wider perspective requires an explanation for the emergence of social norms within primary care. Here, Elster (1989) and Naylor (1990) provide a starting point for understanding the conditions under which status competition between GPs may arise as a relevant phenomenon.

Economic science has hitherto enquired relatively little into the field of non-financial incentives. The scant literature reviewed in this chapter suggests that we should expect intrinsic and social incentives to be of importance. Moreover, we should recognise that they are likely to interact with financial and regulatory incentives, with a danger of neutralising them. In this regard, there is a more general case for exploring the role of organisational culture as a determinant of GPs' quality incentives (Kreps 1990; Davies et al. 2000). Finally, the interaction between the various incentives is complex and highly sensitive to psychological and sociological factors. There remains broad scope for interdisciplinary research to develop a more precise picture of the interaction between psychological, sociological and economic factors.

Summary

- Physicians' altruistic concerns about their patients' wellbeing mitigate the potential for under-provision of quality; but usually not the potential for inefficient resource use. In this case, cost-sharing can induce altruistic physicians to provide optimal levels of quality.
- Intrinsic motivation is likely to be an important quality incentive for physicians. It can be undermined by both market and regulatory incentives.
- Social interaction is another likely source of quality incentives. Competition for social status within a peer group or within society in general may provide quality incentives but also exposes physicians to a 'status' risk that should be accounted for in the design of formal performance schemes.
- Generally, the inter-relationship between the non-financial quality incentives arising from altruism, intrinsic motivation and social interaction, and financial quality incentives is under-researched both from a theoretical and an empirical view.

8 SELF-REGULATION AND CLINICAL GOVERNANCE

8.1 Collective reputation

As we have seen, lack of information on the part of the regulator and the possible crowding out of physicians' intrinsic motivation place formidable constraints on the regulator's ability to control quality provision. These difficulties are reflected in the extensive reliance on self-regulation by the clinical profession in most health care systems. However, in recent years concerns have been growing in many countries that professional self-regulation is not as effective and efficient as it should be. This concerns both the profession's failure to play its part in controlling health care expenditure and its failure to safeguard patients against physicians' poor quality.

In principle, one would expect a profession to have an interest in effective self-regulation in as far as this serves to maintain a collective reputation for quality and allows the profession to charge a quality premium for its services. Collective reputation being subject to a strong free-rider problem, its maintenance requires an effective monitoring and enforcement mechanism. Why does self-regulation appear to fail on occasions?

Example: Within the UK NHS there is considerable concern about the ability of current systems of self-regulation in safeguarding good performance by GPs (Department of Health 1999). One of the reasons is the GPs' status as independent contractors, such that the only two effective way of removing delinquent physicians from the NHS is via NHS Tribunals or by their removal from the Medical Register by the General Medical Council (GMC). Both processes are usually lengthy and, since the physicians concerned are allowed to continue to practice in the meantime, an issue arises about the protection of patients. Finally, there are also difficulties caused by the need to exchange information between NHS bodies and the GMC.

Surprisingly, while it may work at the level of individual physicians, the reputation mechanism may break down for the profession as a whole. I argued earlier that reputation becomes a guarantor of quality

only if patients are able to punish low quality GPs by switching away from them. This requires that patients are able to seek care from an alternative doctor. Consider, in contrast, a collective of GPs trying to maintain a collective reputation. The problem is that the gains from reputation are much lower for the collective than they may be for an individual physician. To see this, suppose the collective, as a whole, offers low quality. In order to punish them, a patient would have to find an outside provider. This is obviously more difficult as now a whole group is under-providing quality, and it may be impossible if this group constitutes the entire profession. In that case, the group of GPs as a whole can afford to not be particularly concerned about its collective reputation (at least not for financial reasons). This in turn curbs their incentive to establish effective self-regulation.[62]

8.2 Cartelisation

Furthermore, some commentators argue that professionals use self-regulation as a device for collusion. Zweifel and Eichenberger (1992) identify a number of institutional factors that facilitate cartelisation. These include:

- universal health insurance coverage, perhaps including a redistributional function;
- presence of fixed fee schedules, determined by bilateral bargaining;
- significant levels of political lobbying and logrolling[63], as measured by the share of health care spending in total public expenditure; and
- a low degree of foreign competition.

62 I have also argued in section 4.3 how collective reputation may fail as a guarantor of quality due to a free-rider problem in which an individual GP faces an incentive to offer low quality while still benefiting from the group's good reputation. Conversely, if the group's reputation is poor, then the efforts of an individual GP to raise quality may not be rewarded.

63 Logrolling is the exchange of political favours between policy makers and/or pressure groups. For example, an offer by the profession to subject itself to self-regulation against a commitment by the policy maker not to engage in external regulation can be viewed as a form of logrolling. The policy maker gains a reputation by being able to announce the profession's vow to self-regulate as an economic alternative to costly intervention, while the profession gains from the absence of external intervention.

Accreditation can be used to control entry into the market and, thus, physician density. The issuance of medical guidelines and treatment standards allows the foreclosure of the primary care market to outside providers offering alternative medicine.

Zweifel and Eichenberger (1992) argue that the respective weights of these factors suggest the following ranking of countries with respect to the degree of cartelisation present: UK and Switzerland > France and Germany > US, Belgium and Sweden. With the exception of the US, this ranking is confirmed by the time trend of physician density (1965-1980), where the increase is the least pronounced in the UK and Switzerland and most pronounced in Belgium and Sweden. Similarly, the trend in income is strongly negative in Belgium and Sweden; in France and the US there was a weak negative correlation between income and physician density, whereas in Germany and Switzerland physicians were able to maintain their income relative to the average.[64]

While this evidence suggests that medical markets are cartelised to significant degree, it should be borne in mind that quality premia, which are necessary for reputation to function as a guarantor of quality, are only secured if physicians have market power in some sense. Recall from section 4.3, that if physicians could not guarantee themselves a fee in excess of their marginal cost, they would have no incentive to maintain a reputation by providing a quality service on a continuous basis. In the light of this, the bias of self-regulation towards cartelisation may bear an indirect benefit at least in terms of maintaining quality.[65]

8.3 Clinical governance

Under the general stance that self-regulation leads to an insufficient control of quality, health care administrators now seek new means of striking a workable balance between external intervention and the professions' self-regulation.

64 In the UK, physician incomes were uncorrelated with physician density. Variation in that period was mainly for political reasons.
65 See Scarpa (1999) for a survey of self-regulation in a variety of markets.

Example: Within the NHS in England, the implementation of national performance standards is left to lower-tier organisations, in particular Primary Care Trusts (Department of Health 1998, 1999, 2000a). It is expected that they implement comprehensive systems of clinical governance to improve care across the GP practices in their area (see Figure 1.1 in section 1.1 above). Crucial elements of clinical governance include:

- *a comprehensive programme of quality improvement activity (such as clinical audit and evidence based-practice) and processes for monitoring clinical care using effective information;*
- *clear policies aimed at managing risk, including procedures that support professional staff in identifying and tackling poor performance;*
- *clear lines of responsibility and accountability for the overall quality of clinical care (Department of Health 1998).*

This is supplemented by a framework for continuing professional development. While the implementation of standards is left to the Primary Care Trusts, they are accountable to the Department of Health and subject to audit by the Commission for Healthcare Audit and Inspection.

Campbell et al. (2001c) point out that clinical governance requires Primary Care Trusts to develop a form of corporate culture, involving shared learning, open exchange of information, the setting of performance targets together with monitoring and, if necessary, enforcement of achievement. With the concept of clinical governance still developing, let me in the following paragraphs review a number of issues emerging from economic reasoning.

First, in contrast to a regime of 'pure' professional self-regulation with no governmental oversight, clinical governance is embedded within a framework of external regulation. With a Primary Care Trust being accountable to the national regulator, the new regime can be viewed as a form of hierarchical regulation, with the implementation of the regulatory framework being delegated to a lower tier. In this regard, the national regulator has to strike a balance between inferior information about the details of local practice and patient data, on the

one hand, and superior enforcement power and ability to co-ordinate regulation across different Primary Care Trusts, on the other. Caillaud et al. (1996) provide a framework in which to analyse the optimal degree of delegation taking into account these factors.

Second, if clinical governance is understood as a set of formal and informal rules it can be interpreted as a form of corporate culture. Kreps (1990) argues that, if chosen appropriately, the principles and rules underlying corporate culture can serve as a focal point for the behaviour that should be expected under unforeseen contingencies. In this regard, it allows the development of reputation even under substantial uncertainty.[66] Here, reputation works in the favour of more than one actor. By sticking to the rules, physicians can establish a reputation both with patients and with the regulator. Patients benefit from having a guarantee of at least some level of quality, while the Primary Care Trust benefits from a lower cost of monitoring GPs. Furthermore, by establishing a framework of clinical governance, the Primary Care Trust itself can establish a reputation with the constituent GPs and with the national regulator. GPs who comply with the rules are protected against discretionary regulatory interventions and attempts by the payer to reduce their income. Individual physicians are, thus, more likely to join the profession and invest in medical skills. The national regulator benefits from a lower monitoring cost.

Clinical governance may thus allow some control of quality while at the same time preserving mutual trust between the various participants in the complex production process of health care. Notably, this is possible even if a formal regulatory and enforcement framework is ruled out and even if quality is unobservable or unverifiable. The caveat is that the framework of rules that constitutes clinical governance must fulfil the following three requirements: (i) compliance with the rules must be observable by the relevant actors;

66 Huntington et al. (2000) and Rosen (2000) point out that implementing clinical governance within a primary care context is likely to be problematic due to the great variability in practices and practice styles and due to the absence of clear hierarchical structures. This corresponds to a significant degree of uncertainty within the organisation.

(ii) the rules must be flexible enough to be relevant even under unforeseen circumstances; and (iii) the rules must reflect true quality and must not give rise to distortions in unregulated dimensions of quality or in efficiency. Drawing up a set of rules that does not compromise on any one of these criteria is a formidable task.

Third, collective learning and information sharing are aspects of clinical governance that go well beyond the task of regulation. Indeed, clinical governance is expected to help to introduce a sort of team spirit amongst primary care professionals, who may hitherto have worked in isolation. From a theoretical perspective, this embraces the form of social interaction (in the form of quality circles, etc.), which we characterised in section 7.3 as one possible mechanism behind the provision of quality. Furthermore, information sharing and collective learning are likely sources of beneficial knowledge spill-overs between GPs. Many commentators on clinical governance stress its participatory character. If physicians are involved in setting their own targets, this is more likely to reconcile the regulatory aspect of clinical governance with the agents' intrinsic motivation. As the discussion in section 7.2 has shown, avoiding a crowding-out of intrinsic motivation could give substantial leverage to clinical governance even if the formal quality incentives are weaker.

There is some empirical evidence to lend support to this view. Campbell et al. (2001b) find that a good team climate is positively correlated with higher outcome scores in various quality dimensions (quality of diabetes care, access, continuity of care, and interpersonal care). Encinosa et al. (1997) have estimated the effects of financial incentives on the average number of consultations between physicians per day. Their results indicate that strong financial incentives curb the physicians' propensity to exchange information.

Whereas a form of participatory regulation suggests strong advantages over direct regulation, there is a major caveat to it. As Campbell et al. (2001c) point out, much of the current enthusiasm for clinical governance arises from the circumstance that the participatory aspects are much in the forefront during the development and implementation phase. A tension between participation and team-building on the one hand and the actual enforcement of performance

targets is likely to arise, however, once Primary Care Trusts are fully accountable for quality. The current supportive mood amongst practitioners might wane once enforcement sets in to bring non-compliant practices into line.

Furthermore, an incentive system relying on the participation of the agents provides scope for the latter to influence the regulator in order to further their own position (Milgrom and Roberts 1992, chapter 8). The introduction of formal incentives usually implies financial bonuses for good performance or non-monetary rewards in the form of increased status. GPs may seek to attain these rewards not just by improving performance but also by influencing the decision-maker. Influencing activity includes the provision of distorted information, lobbying activities and coalition building in the case of collective decision making.

These activities are more pronounced the more open is the decision-maker to communication from the GPs, and they cause two forms of inefficiency. The decisions themselves are likely to be distorted. But, in addition to this, influencing activity is unproductive as it is merely directed at a redistribution of income or other rewards, and the effort so expended therefore constitutes a waste of resources. After all, the time spent by GPs in bargaining for a favourable regime of clinical governance might be better spent on the provision of care.

Summary

- Within many health care systems, concerns are voiced that while the self-regulation of quality by the profession is indispensable it proves to be insufficient. One possible explanation is that fee-riding leads to a lack of incentives to maintain a collective (rather than an individual) reputation.
- Clinical governance has recently received substantial attention by policy-makers and researchers as a potentially powerful mechanism of controlling quality. It is recognised that clinical governance embraces a mix of external regulation and self-regulation.
- In England, Primary Care Trusts are expected to implement national performance standards by introducing a system of clinical governance. In this regard, the bodies within the Primary Care Trusts who are responsible for clinical governance take on a function as supervisors in a hierarchical agency.
- Clinical governance can be understood as a framework of simple formal and informal rules for appropriate behaviour under unforeseen contingencies. In this regard it facilitates the establishment of reputation (by physicians and by the regulator alike) in a complex environment.
- Collective learning and information sharing are understood to be key elements of clinical governance. They can be viewed as a form of participatory regulation, where physicians are involved in determining their own performance framework. While this facilitates the regulator's task, it opens a channel for possibly wasteful (from a societal perspective) influencing activity.

9 ORGANISATION OF THE PRIMARY CARE SECTOR: IMPLICATIONS FOR QUALITY

So far, we have circumvented the quality implications of the organisational form that primary care takes. Organisations have the horizontal dimensions of scale and scope as well as the vertical dimension of integration (Milgrom and Roberts 1992, chapter 16). 'Scale' may be measured, for instance, by the population served by a primary care practice or by the number of GPs working within it. 'Scope' refers to the (horizontal) range of activities, i.e. the range of primary care services offered plus, perhaps community health services or alternative medical services. Finally, the degree of vertical integration measures the extent to which the organisation operates in different stages of production, e.g. the extent to which primary care practices offer specialist services that traditionally belong in the domain of secondary care.

Data on the structure of European primary care organisation reveals wide differences between countries (e.g. Fleming 1992). The following trends are prevalent, however. In those health care systems without a strong tradition in General Practice (e.g. Germany and France), physicians tend to work single-handedly and, thus, at small scale. However, to the extent that they are specialised, they tend to be vertically integrated into secondary care. In contrast, those health care systems with a strong emphasis on GPs as gatekeepers (e.g. the UK, Denmark and Norway), practice partnerships are much more prevalent. The degree of vertical integration is very low in these systems. In all European systems, there appears to be a trend towards more and larger practice partnerships. This is likely to go hand in hand with an increase in the scope of primary care activity.

The purpose of this chapter is not to explain the particular patterns of organisation of health care systems but merely to explore some of the possible consequences of (re-)organisation for quality provision.[67] I consider in turn the horizontal dimensions of scale and scope and then

67 For an introduction to the organisation of business firms, see Milgrom and Roberts (1992, chapter 16). For an application of transaction cost theory to the reorganisation of the NHS see Croxson (1999).

some issues relating to the role of primary care in the vertical structure of the health care system.[68]

9.1 Horizontal structure: quality in group practice

The size of primary care practices, as measured by the number of participating GPs or by patient list size, varies significantly across and within countries. From a theoretical perspective, the concept of (dis-) economies of scale can be conveniently employed as a way to predict the optimal size of firms or practices.[69] Economies of scale are present as long as further increases in activity lead to a reduction in average cost. Common sources of economies of scale are:

- fixed costs, arising from investment in equipment or from fixed labour contracts;
- increasing bargaining power vis-‡-vis suppliers;
- learning effects; and
- knowledge spill-overs among a greater number of staff.

For most production processes, there exists a level of activity beyond which diseconomies of scale set in, i.e. a level beyond which greater activity leads to rising average cost. This is the case if the organisation works close to full capacity and limiting factors, such as management resources, technical equipment or the communication infrastructure, give rise to bottlenecks.

If economies of scale arise for low levels of activity but turn into diseconomies at high levels, there exists an optimal size of an organisation at which average cost is minimised. A substantial amount of empirical work has been carried out investigating the optimal size of health care institutions.[70] This research is mostly couched in terms of costs and resource use. However, as we will see presently, there are also implications for the provision of quality.

68 The issues involved parallel those in the US-context of managed care. In order to limit the scope of this overview, however, I do not discuss this US literature and merely refer the reader to Glied (2000).
69 See Cabral (2000, chapter 2) or Tirole (1988) for an introduction to these concepts.
70 A current debate in the UK focuses on the optimal size of Primary Care Trusts (Bojke et al. 2001).

Research into professional partnerships has established an inter-relationship between the sharing of income risk, the strength of internal incentive systems, i.e. performance pay, and the size of the partnership (Gaynor and Pauly 1990; Gaynor and Gertler 1995; Encinosa et al. 1997). Risk-averse agents benefit from partnerships, as they are able to pool their income risk by way of sharing profits and losses. On the downside, profit sharing gives each member of the group an incentive to under-provide effort as financial rewards are now less sensitive to personal effort. The relationship between group size and incentives is, therefore, not straightforward. For any given degree of profit sharing, an increase of group size entails a greater incentive to under-provide effort. This is because the greater the number of contributors to the shared profit, the less sensitive it is to any individual's effort. But, for the same reason, a greater group also implies a reduction in the risk faced by an individual physician. The lower risk allows the group to reduce the degree of profit sharing and, thereby, enhance the strength of incentives. One would thus expect group size and the degree of profit sharing to be determined jointly.

Gaynor and Gertler (1995) demonstrate this empirically and identify risk aversion as the driving force. Furthermore, their model allows for a direct interpretation in quality terms. They assume that patient demand for consultations with a physician increases the better is the quality of service, where quality is observed by patients but not the econometric researcher. Quality itself is a function of the physician's effort, a set of non-physician inputs, practice capital, and the physician's skills. With the other variables being unrelated to the compensation system, at least in the short run, the key driver behind effort and, thus, quality turns out to be the degree of profit sharing. Gaynor and Gertler (1995) show econometrically that partnerships between risk-averse physicians are characterised by a greater degree of profit sharing and a smaller size, the latter being a response to the lower incentive to provide effort. They also show that the effect of profit incentives on quality and, thus, income dominates the effect of group size such that risk-averse partnerships tend to produce lower quality and generate less income.[71]

As a caveat, it should be noted that Gaynor and Gertler's (1995)

empirical findings provide no direct evidence of the effects of risk aversion and partnership size on the quality of care. While the above interpretation is suggested by the way they have set up their model, the following interpretation suggests a negative effect of risk aversion on quality in spite of being equally consistent with their results. Suppose that demand, i.e. the number of consultations, is determined by physicians, who face a constraint regarding their total working hours. Then, a lower degree of profit sharing provides incentives for the physician to increase the number of consultations (and services). One way of achieving this is by reducing quality, e.g. by curbing the time spent on diagnosis or giving advice to the patient.[72] Hence, while being more productive in terms of income generated, practices with lower degrees of risk aversion provide lower quality. We see that once again the impact of incentives on quality depends crucially on whether or not patient demand is sensitive to quality. Unfortunately, this relationship has remained mostly unidentified in empirical terms.

Encinosa et al. (1997) demonstrate that the findings by Gaynor and Gertler (1995) must be modified if status competition plays an important role within the partnership. In this case, incentive systems are more likely to be implemented in small groups. If status increases with income, then income risk is now complemented by the risk of losing status. The loss of status within the group is the more pronounced the greater is the number of partners. But this implies that a larger group does not necessarily imply a lower risk for the individual. While a larger group reduces the income risk, it exacerbates the potential loss of status. As Encinosa et al. (1997) suggest, this leads large partnerships to reduce the power of their incentive systems, e.g. by introducing a greater degree of profit sharing.

The overall effect on productivity and quality are, therefore, even more blurred. On the one hand, status competition may rise to a direct incentive for the provision of quality, in particular, if status is

71 The income foregone can be interpreted as an insurance premium for risk reduction, which increases with the degree of risk aversion.

72 Note that this relates to the multi-task problem raised in section 6.1. In the present interpretation, reward is linked to the quantity of the physician's output rather than to its quality.

not only determined by income but also by clinical success or by popularity amongst patients. Here, status competition and the ensuing incentives for the provision of quality are the more pronounced the larger the group. On the other hand, as we have just argued, the degree of profit sharing is likely to be greater in large groups, with an unclear effect on quality. The picture that emerges must therefore remain sketchy. Whereas there is agreement on the importance of the interaction between risk, incentives and group size, the contradictory findings hitherto strongly suggest a need for further research.

The greater extent of knowledge spill-overs within partnerships, i.e. the exchange of professional experience and skills, is likely to enhance quality. This leaves open the question as to what might be the optimal size for such an exchange of professional expertise. Initially, spill-overs are likely to increase with the number of doctors, but very large practices may stifle professional communication due to an atmosphere of anonymity. Furthermore, spill-overs are likely to be significant only if GPs are willing to share their knowledge with partners. As we have already discussed in the context of clinical governance, this requires a degree of co-operation as opposed to rivalry within the organisation. Encinosa et al. (1997) have estimated the effects of group size and the strengths of incentives on the average number of consultations between physicians per day. They find that while the propensity towards professional exchange increases with group size, the effect of strong monetary incentives is negative.

Greater scale may allow primary care physicians to specialise on certain aspects of care. Furthermore, larger practices are better able to employ nurses and managers. Specialisation sets free physicians' time and allows each 'specialist' to realise a greater degree of 'learning by doing' in the chosen activity. Both of these factors are likely to contribute to greater quality. These theoretical arguments notwithstanding, a recent study by Hippisley-Cox et al. (2001) could not confirm significant differences in performance between single-handed and group practices once the lower average socio-economic status of patients attending single-handed practices was accounted for. Campbell et al. (2001b) showed that while larger practices performed better in some aspects of clinical quality (diabetes care), they were

outperformed by smaller practices in terms of ease of patient access.

An argument that is conceptually similar to the one for 'scale' can be made with regard to 'scope'. Economies of scope exist if the cost of producing a range of services together is lower than the cost of producing them separately. Economies of scope can arise due to shared inputs, such as management or common infrastructure, or due to the diversification of risk. More precisely, partners can reduce their income risk if they can add services that generate income, which is unrelated, or even negatively related, to the income from the established range of services on offer.[73] Profit-sharing only insures a physician against shocks affecting individual income but not against shocks affecting the group's income. In contrast, diversification of services insures the whole group against variations in income. For example, by taking on board services from the field of alternative medicine a partnership can insure itself against a shift away from traditional medicine. As the lower risk within a well-diversified partnership allows its members to operate under a lesser degree of profit-sharing, stronger incentives for efficiency and quality can result. Risk aspects aside, the provision of a greater range of services within a single practice may itself be viewed as an improvement in quality.

Policy implications: The horizontal organisation of primary care is not only of interest at practice level. The Primary Care Trusts in the UK comprise a number of individual GP practices and can, thus, be viewed as analogous to firms operating a number of plants. The range of relevant organisational issues include the resource allocation processes within the organisation (Dusheiko et al. 2001) and the determinants of optimal size (Bojke et al. 2001). Their implications for quality remain to be explored.

73 The insurance argument does not apply if the partners specialise in complementary services. In this case, lower demand for service A entails a lower demand for service B as well and, therefore, implies an even greater income risk.

9.2 Vertical structure: GP as intermediary in the production of care

I stressed earlier that the production of health care can be viewed as a two-stage process, in which secondary care, primary care and pharmaceutical inputs are produced at stage 1, and are then combined by primary care physicians at stage 2. When addressing the vertical organisation of production, researchers confront two broad sets of questions.

First, taking as given the separation of actors at different stages, how do manufacturers of inputs interact with the assembler or retailer of a final product? Inter alia, this involves issues about the composition and promotion of the final product on the part of assemblers/retailers; their incentives to control the quality of inputs and the effect of this on the manufacturing decisions; and the forms of contractual arrangements governing the process (Cabral 2000, chapter 11).

The second set of questions relates to how far the process of production should be vertically integrated, i.e. carried out within one organisation (Milgrom and Roberts 1992, chapter 16). The answers to these questions usually relate to the following determinants: (i) the relative costs of organising production, i.e. the transaction costs of writing and enforcing contracts between independent parties versus the costs of governing a hierarchy within an organisation; and (ii) the degree of market power that the firm(s) in question command or acquire under the respective organisational structures.

Alternatively, the issue of vertical organisation can be approached within the framework of hierarchical agency (Tirole 1986, 1994; Caillaud et al. 1996). Here, the simple model involving the delegation of a task from a principal to an agent is extended to allow for at least one intermediate stage of supervision. The principal enters into contracts with a producing agent and with a supervisor. In this, the principal either relies on the supervisor's direct control of the producing agent or on the supervisor's reports about the agent's performance. Applying these ideas to the context of health care, one could envisage the following hierarchy. The payer controls two groups of agents: hospitals as providers of secondary care and GPs as providers

of primary care. GPs act as intermediaries in assembling the overall bundle of care but have an additional role in auditing the quality of secondary care.

Modelling the vertical process is a demanding task, not least because in most cases there are elements of both 'competition' and 'hierarchy'.[74] Given the complexity of the problem, which to my knowledge has not yet been addressed adequately, let me point out a number of issues, which are likely to be of some relevance in the context of vertical relationships.

9.2.1 The GP as commissioner of secondary care

Let us first consider the role of the GP as an assembler of care when the boundary between primary and secondary care is given. The issue is not one of vertical integration, but rather of how GPs assemble care and then whether they function as an effective auditor of the quality of secondary care. Furthermore, it may be asked which contractual arrangements optimise the overall quality of care.

As discussed earlier, the physician payment system is one key determinant of the input mix chosen by the GP in the assembly of health care. More specifically, capitation and salary lead to an incentive to increase onward referrals to specialists, while we have found the reverse to be true for FFS and fixed budgets. Furthermore, if patients equate specialist care with high quality, then use of specialist secondary care inputs increases with the degree of competition between GPs. Finally, the risk to which a GP is exposed under a particular payment or regulatory regime bears on the structure of care. The budgetary risk under fundholding may induce a GP to under-refer patients, whereas the risk arising from malpractice litigation or from quality related performance pay may induce excessive use of diagnostic and other services.

All of these determinants relate to incentives arising within the primary care stage. One should expect, however, that the quality and

74 The relevance of the interplay between markets and hierarchical organisation has been addressed in an informal way in the theory of quasi-markets. See, for example, Bartlett and LeGrand (1993) and Propper (1993).

– where applicable – the price of secondary care and pharmaceutical services play an equally important role in determining the mix of services that GPs assemble for their patients. A full understanding of the composition of care at the primary care stage and its implications for quality, therefore, requires the recognition of the market and/or regulatory conditions in the secondary care and pharmaceutical segments. The interaction between primary and secondary care providers and the impact of regulatory incentives in such a system is complex. Regulation in one of the segments is likely to have an impact on behaviour in the other segment. As one illustrative example of the interaction between primary and secondary care, I focus in the following on the role of GPs in guaranteeing the quality of secondary care.

It has been argued that, by specialising in diagnosis, gate-keeping GPs can act as credible advisers on the composition of care and as guarantors of the quality of secondary care (Dranove and Satterthwaite 2000). Similar to retailers in other sectors, or to financial intermediaries, they are fit to judge the quality of secondary care services and, under appropriate incentives, 'market' to patients only services of sufficient quality.[75] It is, thus, hoped that the selection of quality services by GPs induces a form of quality competition amongst the providers of secondary care. However, this is subject to a number of caveats.

First, diagnosis and the evaluation of secondary care quality require effort by the GP and, therefore, appropriate reimbursement. Second, impartiality in diagnosis and judgement of quality requires that GPs do not take a personal interest in the composition of care.

75 The role of intermediaries has been addressed in the industrial organisation literature. Biglaiser and Friedman (1994), for example, expand their reputation and signalling model (see section 4.3 above) to include intermediaries who market a variety of products they purchase from 'single good' manufacturers. When selling a low quality product, an intermediary loses reputation across the board of all the products offered. Therefore, intermediaries face a greater loss from passing on low quality goods or services than manufacturers do and, therefore, a stronger incentive to maintain a good reputation. The societal cost under which reputation guarantees the provision of quality is, thus, lowered by the presence of intermediaries. Again, it would be instructive to apply this model to the health care context.

But, as we have seen, the payment system and the presence of risk are likely to distort referral decisions. In particular, the extent to which GPs are exposed to the financial consequences of referrals is likely to have significant effects on the quality of secondary care. If GPs do not have to bear the cost of referrals, their only concern lies with quality. Secondary care providers are likely to respond to that, resulting in quality competition between hospitals which may induce them to provide high, possibly excessive, quality.[76] Furthermore, quality may be distorted according to GPs' rather than patients' preferences. If, in contrast, fundholding GPs have to bear the full cost of referrals, a concern about the cost of secondary care may lead to price rather than quality competition between secondary care providers and this may even be at the expense of quality.[77]

While the issue of quality competition versus price competition in the hospital market has been primarily discussed in the US context (Dranove and White 1994; Dranove and Satterthwaite 2000), the general lessons from this are important for those European health care reforms that are aimed at greater competition between health care providers.

In many cases, the commissioning of secondary care by primary care decision-makers is unlikely to follow a market process. It is more likely to be a matter of bargaining over a contract. Caillaud et al. (1996) consider a model of a three-tier-hierarchy, which could be used to address the question as to what extent the delegation of commissioning power to primary care institutions can enhance the efficiency and quality of service provision. In this model, the principal determines the degree of authority granted to an intermediate agent who then bargains with a lowest level agent about a contract specifying a task to be carried out together with a payment. This reflects the decision a health authority takes about the delegation to primary care institutions of authority in purchasing and controlling secondary care.

76 This has been found for the US before the advent of prospective hospital payment and intense price competition driven by cost-conscious health plans (Dranove and White 1994; Dranove and Satterthwaite 2000).

77 See Dranove and Satterthwaite (2000) for a discussion of these, more recent, concerns in the US health care system.

Caillaud et al. (1996) show that the optimal degree of delegation is attained by balancing the informational advantage held by the intermediate agent, who is better informed about the characteristics of the lowest level agent, against the greater bargaining power held by the principal. Accordingly, more authority should be granted to primary care institutions the greater the health authority's deficiency of information about patient needs and the quality and cost of secondary care, and the greater is the primary care institution's bargaining and enforcement power. In this regard, it is a matter for debate whether the UK policy of granting Primary Care Trusts strong commissioning responsibilities is appropriate. Whereas Primary Care Trusts are likely to have better access to decentralised information about their patient population than higher level institutions, it is unclear whether they enjoy sufficient bargaining power vis-‡-vis their secondary care suppliers.

Primary care physicians are sometimes envisaged as taking on a role as 'whistleblowers' with respect to poor quality secondary care. Whistle-blowing may be achieved by an explicit reporting system, by informal communication between the regulator and GPs, or by the regulator observing some signal of GP activity such as shifts or intended shifts in referral patterns.

In this regard, GPs have a role akin to supervisors of production within a hierarchy. Tirole (1986) gives an interesting account of the problems that arise due to possible collusion between supervisors and supervisees. Indeed, tacit collusion between physicians is likely to be a major problem in the self-regulation of the profession. In this regard, one could express two opposing views on the role of GPs as (implicit) auditors of secondary care. On the one hand, their role in quality assurance is likely to be hampered by the presence of informal links with secondary care providers.[78] On the other hand, if they have sufficient concern for their patients' welfare, they may serve as better guarantors of quality than disinterested external auditors of secondary care. However, in the case of GPs acting as whistleblowers there may

78 This may be particularly likely in an environment that deliberately fosters co-operation rather than competition (Goddard and Mannion 1998).

be a danger of GPs talking down the quality of secondary care if this helps them in improving the regulator's perception of their own contribution to the provision of health care.

9.2.2 Prescription

Similar issues arise from the GP's role in prescribing medicines. Again, physicians function as agents to both the patient and the payer. In this regard, they face incentives in allocating medicines to patients on grounds of their appropriateness and effectiveness as well as on grounds of their cost-effectiveness. However, the agency problem is more complicated due to the interaction between pharmaceutical companies and the physician.

Producers of medicines undertake considerable 'detailing' efforts in providing information to physicians on their existing and new drugs. While there is no doubt about the benefits from detailing in improving the information of physicians with regard to the attributes of certain drugs, concerns are sometimes voiced that detailing may also – and perhaps predominantly – function as a form of persuasive advertising. As a consequence, physicians may be induced to prescribe a particular medicine to patients despite the existence of superior substitutes or despite the ineffectiveness of the medicine. Both incentives imply a sub-optimal provision of quality either directly in the form of allocative inefficiency (distorted prescription) or indirectly in the form of productive inefficiency (care not produced at minimum cost).

Recent attempts have been made to model the physician as a triple agent of the patient, payer and the pharmaceutical producer. Konrad (2002) demonstrates in a theoretical model how the presence of detailing by pharmaceutical companies may distort prescription decisions. Analysing Swedish data, Lundin (2000) provides empirical evidence for the presence of moral hazard on the part of physicians who are less likely to prescribe the more expensive branded drugs when the patient faces a high co-payment.[79]

79 See Goodwin (1998) for some rather mixed evidence of the effects of fundholding on prescriptions.

9.2.3 Co-ordinating primary and secondary care

Recent years have witnessed a discussion about the optimal balance between primary and secondary care (Scott 1996; Godber et al. 1997; Saltman and Figueras 1997, chapter 6; Dixon et al. 1998). It has been placed against the need to control health care expenditure, on the one hand, and the recognition that by assembling health care, primary care physicians exercise significant control over health care resources. At the core of this discussion is the extent to which primary and secondary care should be integrated in order to guarantee efficient resource use and the provision of quality by way of better co-ordination.

A possible lack of co-ordination between primary and secondary care has been diagnosed by a number of researchers (Wilkin and Dornan 1990; Jankowski 2001). In many cases, this relates to a lack of communication. At a process level, this has obvious quality implications, such as longer waiting times due to patients' medical records being lost in the referral process; or duplication of diagnostic tests; or the pursuit of wrong treatment paths. At a structural level, a lack of co-ordination may lead to investment decisions both at primary and secondary care level that lead to a sub-optimal division between primary and secondary care.

Let me consider this last issue within a model of managerial co-ordination devised by Milgrom and Roberts (1992, pp. 111-113). Consider a condition, which can be treated by a combination of primary and secondary care interventions. For instance, cancer treatment may be based on self-medication supported by regular check-ups by a GP and/or radiation therapy within secondary care. Both forms of treatment require an investment in skills and equipment on the part of the respective physician. Suppose initially that both physicians are benevolent and choose their investment with a view to optimising the quality of care for the patient, subject to a resource constraint, but that the decision-makers are unable to communicate with one another.

GPs condition their investment in skills and diagnostic devices on the level of investment they expect in secondary care. For any level of investment within secondary care, the *pc* schedule in Figure 9.1

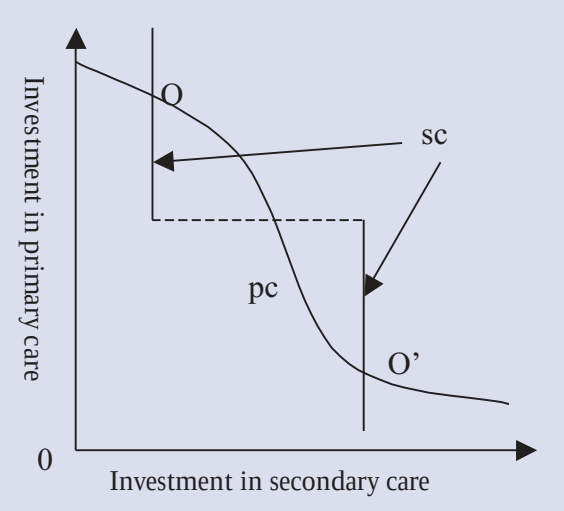
describes how much the GP has to invest in order to maximise patient benefit subject to the resource constraint. If the GP also invests in other activities, the resource constraint gives rise to opportunity costs that prevent investment at the maximum level. Furthermore, suppose that primary and secondary care investments are substitutive in the sense that high levels of secondary care investment imply low additional returns in terms of patient benefit from investment in primary care. Hence, the negative slope of the *pc* curve.[80]

Likewise, secondary care physicians take the level of investment in primary care as given when determining their own investment. This is depicted by the schedule *sc* in Figure 9.1. The *sc* schedule consists of two segments corresponding to high and low investment, respectively. This reflects a discrete decision on whether or not to invest in a piece of equipment, such as a radiation therapy device. Only if the expected level of primary care investment is low, will the investment be made and secondary care be administered at a high level (in the neighbourhood of O'). If the secondary care physician expects a high level of primary care skills, the investment is not undertaken, resulting in a low level of secondary care activity (in the neighbourhood of O).

There exist two congruent patterns of treatment, and implicitly investment, in the sense that primary care and secondary care physicians agree on the investment schedules. In Figure 9.1, the congruent patterns correspond to the points of intersection of the *pc* and *sc* schedules, O and O'. Combination O reflects a primary care intensive treatment, e.g. the case of self-medication under regular checks. Here, the GP undertakes a substantial investment in medical skills and knowledge in order to provide the support required. There is little investment in secondary care, as the radiation therapy device is not purchased. In contrast, treatment at O' is secondary care intensive. Here, the secondary care physician invests in the radiation therapy device and it is then optimal for the GP to refer patients on so that

80 The particular shape of the *pc* curve is determined by the nature of the condition as well as by the way in which primary and secondary care interact in the generation of quality. If investments in primary care and secondary care were complements, rather than substitutes, then the *pc* curve would slope upwards.

Figure 9.1 Co-ordination problem in the composition of care



Source: Adapted from Figure 4.3 of Milgrom and Roberts 1992.

extensive medical knowledge on the part of the GP is not warranted and primary care investment remains low.

A priori, it is unclear which of the congruent treatments is socially optimal. Depending on the type of condition, the physicians' skills and relative costs, it may be either of the patterns at O and O'. The problem is that, no matter which of them is optimal, there is no guarantee that physicians will agree on it. Even worse, once a sub-optimal but congruent combination has been established it is difficult to move away from it. Suppose, for example, that the current secondary care intensive treatment O' is sub-optimal and that a primary care intensive treatment O would give rise to a superior outcome. The problem is that a shift towards the treatment O can only be achieved by a 'jump' in investment levels, which requires co-

ordination. This is because any small deviation from O' even in the right direction gives rise to a poorer outcome in welfare terms than O'. This implies that active co-ordination is likely to be necessary even if physicians are assumed to act in the patients' best interests.

Figure 9.2 Technical progress and design changes

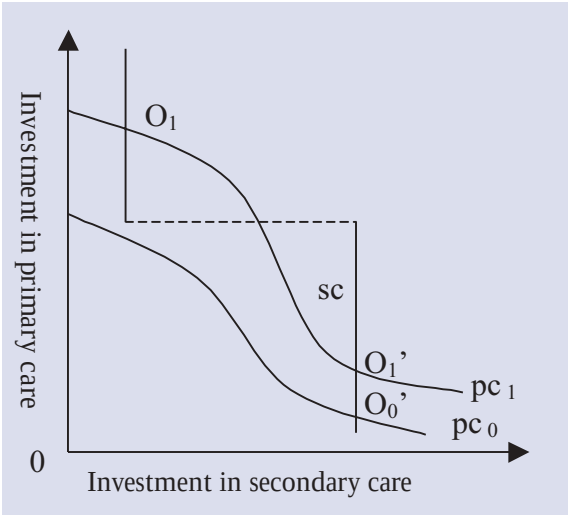


Figure 9.2 shows that a system without integrated decision making may fail to react optimally to technological change. This is particularly likely, as the two physicians frame the information about new technologies in the context of their experience. Consider, for example, the development of an effective drug, the administration of which requires close scrutiny by a skilled primary care physician. This corresponds to an outward shift of the primary care reaction curve from pc_0 to pc_1 in Figure 9.2.

Before the innovation, a secondary care intensive treatment at O_0' was the only one congruent and was, indeed, socially optimal. After

the innovation, there are two congruent combinations of treatment: at O_1 and O_1'. Suppose the innovation has altered outcomes such that treatment at O_1 is optimal now. It is likely that, in the absence of co-ordination, the doctors will fail to reach the investment levels corresponding to treatment O_1. In response to the innovation, GPs are likely to engage in a tentative expansion of the primary care surveillance and continue to refer patients. The secondary care sector is likely to continue with investing in the radiation therapy device. The resulting adjustment along the curve terminates at congruent treatment pattern O_1', which implies only a modest change in the balance of primary and secondary care and a sub-optimal outcome as compared to O_1. Without a co-ordinated assessment of the wider-ranging implications of technological change, a major switch in the treatment pattern cannot be accomplished.

Policy implications: The example above underlines concerns about the co-ordination of primary and secondary care and, more generally, between health and social care (Department of Health 2000a). There are two types of options for achieving this co-ordination:

- *deepening the degree of vertical integration of primary and secondary care decision-making. This can be achieved either by allocating the decision rights to a single actor (e.g. the Primary Care Trust); or by implementing a process of explicit co-ordination and joint decision making. The importance of the latter approach is emphasised by the growing importance of networks in health care (Eastham and Ferguson 2003);*
- *central intervention. This requires the specification and enforcement of explicit guidelines about the structure of care. This is the policy approach embraced by the UK NHS's National Service Frameworks (Department of Health 1998, 1999, 2000a). Here, clinical pathways for key conditions, as well as the implied responsibilities of primary and secondary care organisations, are specified in considerable detail (e.g. Department of Health 2000b for Coronary Heart Disease). These guidelines imply a certain structure of care and the investment choices that implement it.*

Summary

- The horizontal organisation (scale and scope) of primary care and its position in the vertical production chain of health care provide important quality incentives.
- Practice size enhances quality as larger practices provide greater scope for professional exchange. On the other hand, if practice partnerships engage in profit sharing, incentives to free-ride tend to increase with the number of partners. This may stifle quality incentives. Practice size and the degree of profit sharing are likely to be determined as a function of factors such as the partners' degree of risk aversion or their social attitudes. The overall implications for quality are complex and require a case by case analysis.
- GPs play an important role as intermediaries in that they commission secondary care and/or audit its quality. In their role as commissioners they may induce a form of quality competition between providers of secondary care. In their role as auditors they act as intermediate agents to the regulator with a 'whistle blowing' function.
- GPs are also intermediaries in prescribing medicines. Here, the role of information provided by pharmaceutical companies is ambiguous. It helps to inform physicians' decision-making and so improve quality of care, but it might also be used as a form of persuasive advertising and, thereby, distort prescription choices.
- The co-ordination of care at the primary and secondary care interface has profound consequences for quality. Economic modelling can be used to illustrate potential co-ordination problems regarding investment in skills and technology, which may bias treatment patterns towards sub-optimal solutions.10

10 CONCLUDING REMARKS

This work has sought to shed some light on the issue of quality provision in primary care by reviewing the relevant health economics literature and identifying applicable insights from other branches of economics. As expected, the emerging picture is complex and less than clear. However, a number of conclusions can be advanced beyond the (tentative) policy implications that were highlighted throughout the text and shall not be repeated here.

While a workable concept of quality and quality production can be developed for theoretical purposes, it is difficult to apply it in practice. This implies that many of the conjectures arising from theoretical reasoning are not easily subjected to empirical testing. Furthermore, it is unlikely that there will emerge a unified view as to what constitutes high quality care. Open debate about quality is, however, an asset rather than an obstacle, because it continuously puts clinical practice and regulatory intervention to the test. As this review has shown, a range of issues from the clinical and health service research debate on quality lend themselves to integration into economic modelling even if this potential has not yet always been realised. This should reassure those who are rightly seeking greater inter-disciplinary debate between some rather secluded groups of scientists.

Much of the current debate on the importance of primary care revolves around the role that physicians play as double agents, representing both individual patients and society as a whole, and the role they play as assemblers of care. These roles prevail within all institutional settings and are of importance for most policy considerations. A substantial part of policy analysis revolves around the effects of payment systems on physicians' behaviour. The insight that capitation payment tends to give rise to under-provision of services and quality, and FFS to over-provision, is well established theoretically and to some extent confirmed empirically. As we have seen, however, the incentives from payment systems are likely to be modified by the information structure, practice organisation, regulatory arrangements and, not least, by the existence of non-financial incentives: altruism, professionalism and status seeking. Thus, one should expect under any payment system a variety in

behaviour, which may depend on a practitioner's personality more than on anything else.

One important insight for the regulator and/or payer of health care is that the provision of quality incentives is usually not costless. In particular, if the regulator/payer or patient lacks information on physicians' skills or levels of effort, then the acquisition of this information and the design of quality incentives usually carry a direct or indirect welfare cost. In many instances this welfare cost may be so high as to rule out the implementation of effective incentives. Other limitations to regulation arise from risk and risk aversion, agent collusion and the potential crowding out of intrinsic motivation. The message to the policy-maker is, therefore: beware of the unexpected side effects of regulation.

The importance of intrinsic motivation and status seeking suggests that the 'presentation' of an incentive system to physicians may be just as important as its design. Indeed, when deciding on the extent to which the profession should be involved in the design and control of regulation, the policy-maker may face a strong trade-off between reducing the negative effects of 'influencing activity' and maintaining the professionals' intrinsic motivation. The problem is further compounded by the likelihood that professionals with high intrinsic motivation may be severely frustrated by the introduction of regulation, whereas others may only respond to regulatory incentives. Thus, the regulator would ideally have to differentiate the incentive system according to physicians' personalities. The practical impossibility of this underscores the difficulties in regulating primary care.

In many cases the best option for the regulator amounts to employing a set of 'soft' policies, including stimulation of professional exchange, improvement of patients' information, encouragement or requirement of continuous professional education, as well as 'moral rewards' for good practice and 'moral sanctions' for unprofessional behaviour. Understood as an amalgamation of such 'soft' measures in the form of corporate culture, clinical governance may have a significant role to play in guaranteeing quality.

The (re-)organisation of primary care clearly matters for the provision of quality. The horizontal scale and scope of activities influence physicians' behaviour through a variety of channels. These include economies of learning by specialisation, gains from risk-sharing, knowledge spill-overs, and the influence of status competition. Again, the influences are highly sensitive to the specific institutional environment, which is reflected in the ambiguous empirical evidence.

The vertical relationship between primary and secondary care also bears on quality both through the GP's choice of health care mix and through the quality incentives arising from this for secondary care providers. While GPs have an important role to play as explicit or implicit auditors of the quality of secondary care, it remains debatable whether or not they face the proper incentives in this. Economic theory can contribute some insights to the debate on whether primary and secondary care should be more integrated. The issue can be viewed as a co-ordination problem in the delivery of care, where separate primary and secondary care decision-makers are not always co-ordinating on a socially optimal pattern of treatment.

Perhaps the least contentious finding is the vast scope for further economic research into this area. From a theoretical perspective, a number of issues still await proper modelling. For example, little is known about the incentives in FFS systems with global budget caps. Furthermore, some systems, such as the German, have a corporatist structure, in which physicians' and insurers' associations negotiate budgets and the fee structure. Little is known about the incentives in this bargaining process, the outcome, and the incentives arising from this for individual physicians. On a more general note, there is still substantial scope for researching the interaction between a regulator and a profession in a relationship that is more likely to be characterised by influencing activities, bargaining and tacit collusion than by commonplace agency. The need for further research extends to vertical relationships, where little is known about the interaction of incentives in primary and secondary care or the implications for regulatory and institutional design.

Empirical evidence on the issues I have reviewed is scarce and in

many cases inconclusive. The problem of finding good measures for quality is seriously impairing empirical work. However, this also demonstrates the remaining scope for empirical research.

In conclusion, let me express the hope that this review has provided some idea about what economic analysis can – and what it cannot – contribute to the understanding of the incentives behind the delivery of high quality (primary) health care.

REFERENCES

Allen, R. and P. Gertler (1991), 'Regulation and the Provision of Quality to Heterogeneous Consumers: the Case of Prospective Pricing of Medical Services', *Journal of Regulatory Economics* 3, 361-375.

Arrow, K.J. (1963), 'Uncertainty and the Welfare Economics of Medical Care', *American Economic Review* 53, 941-973.

Barkema, H.G. (1995), 'Do Job Executives Work Harder When They Are Monitored?' *Kyklos* 48, 19-42.

Bartlett, W. and J. LeGrand (1993), 'The Theory of Quasi-Markets', in J. LeGrand and W. Bartlett (eds.), *Quasi-Markets and Social Policy*, Basingstoke and London: Macmillan.

Berrow, D., C. Humphrey and J. Hayward (1997), 'Understanding the Relation between Research and Clinical Policy: a Study of Clinicians' Views', *Quality in Health Care* 6, 181-186.

Biglaiser, G. and J.W. Friedman (1994), 'Middlemen as Guarantors of Quality', *International Journal of Industrial Organization* 12, 509-531.

Blomqvist, A. (1991), 'The Doctor as Double Agent: Information Asymmetry, Health Insurance, and Medical Care', *Journal of Health Economics* 10, 411-432.

Bloor, K., A. Maynard and A. Street (2000), 'The Cornerstone of Labour's New NHS: Reforming Primary Care', in P.C. Smith (ed.), *Reforming Markets in Health Care. An Economic Perspective*, Buckingham and Philadelphia: Open University Press.

Blumenthal, D. (1996), 'Quality of Health Care. Part 1: Quality of Care – What is it?' *New England Journal of Medicine* 335, 891-894.

Blumenthal, D. and A.M. Epstein (1996), 'Quality of Health Care. Part 6: The Role of Physicians in the Future of Quality Management', *New England Journal of Medicine* 335, 1328-1331.

Bojke, C., H. Gravelle and D. Wilkin (2001), 'Is Bigger Better for Primary Care Groups and Trusts?' *BMJ* 322, 599-602.

Brook, R.H., E.A. McGlynn and P.D. Cleary (1996), 'Quality of Health Care. Part 2: Measuring Quality of Care', *New England Journal of Medicine* 335, 966-970.

Cabral, L.B.M. (2000), *Introduction to Industrial Organization*, Cambridge, Mass. and London, UK: The MIT Press.

Caillaud, B., B. Jullien and P. Picard (1996), 'Hierarchical Organization and Incentives', *European Economic Review* 40, 687-695.

Campbell, J.L., J. Ramsay and J. Green (2001a), 'Age, Gender, Socioeconomic, and Ethnic Differences in Patients' Assessments of Primary Health Care', *Quality in Health Care* 10, 90-95.

Campbell, S.M., M. Hann, J. Hacker, C. Burns, D. Oliver, A Thapar, N. Mead, D. Gelb Safran, and M.O. Roland (2001b), 'Identifying Predictors of High Quality Care in English General Practice: Observational Study', *BMJ* 323, 1-6.

Campbell, S., M. Roland and D. Wilkin (2001c), 'Improving the Quality of Care through Clinical Governance', *BMJ* 322, 1580-1582.

Chalkley, M. and J.M. Malcomson (1998a), 'Contracting for Health Services with Unmonitored Quality', *Economic Journal* 108, 1093-1110.

Chalkley, M. and J.M. Malcomson (1998b), 'Contracting for Health Services when Patient Demand does not Reflect Quality', *Journal of Health Economics* 17, 1-19.

Chalkley, M. and J.M. Malcomson (2000), 'Government Purchasing of Health Services', in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics (Vol. 1)*, Amsterdam: North-Holland.

Croxson, B. (1999), *Organisational Costs in the New NHS. An Introduction to the Transaction Costs and Internal Costs of Delivering Health Care*, London: Office of Health Economics.

Danzon, P.M. (2000), 'Liability for Medical Malpractice' in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics (Vol. 1)*, Amsterdam: North-Holland.

Davies, H.T.O., S.M. Nutley and R. Mannion (2000), 'Organisational Culture and Quality of Health Care', *Quality in Health Care* 9, 111-119.

Davis, P., B. Gribben, A. Scott and R. Lay-Yee (2000), 'The "Supply Hypothesis" and Medical Practice Variation in Primary Care: Testing Economic and Clinical Models of Inter-practitioner Variation', *Social Science and Medicine* 50, 407-418.

Dawson, D., M. Kuhn, A. Street and P.C. Smith (2001), Regulating Health Service Standards to Reduce Variation: An Economic Analysis, paper presented at the International Health Economics Association conference 2001, York.

Deci, E.L. and R.M. Ryan (1985), *Intrinsic Motivation and Self-Determination in Human Behaviour*, New York: Plenum Press.

Demange, G. and P.Y. Geoffard (2002), Reforming Incentive Schemes under Political Constraints: the Physician Agency, paper presented at CEPR workshop on Health Economics and Public Policy 2002, Bergen.

Department of Health (1998), *A First-class Service. Quality in the New NHS*, London: The Stationery Office.

Department of Health (1999), *Supporting Doctors, Protecting Patients. A Consultation Paper on Preventing, Recognising and Dealing with Poor Clinical Performance of Doctors in the NHS in England*, London: The Stationery Office.

Department of Health (2000a), *The NHS Plan. A Plan for Investment. A Plan for Reform*, London: The Stationery Office.

Department of Health (2000b), *National Service Framework for Coronary Heart Disease*, London: The Stationery Office.

Department of Health (2001a), *Primary Care, General Practice and the NHS Plan; Information for GPs, Nurses, Other Health Professionals and Staff Working in Primary Care in England*, London: The Stationery Office.

Department of Health (2001b), *A Commitment to Quality, a Quest for Excellence. A Statement on Behalf of the Government, the Medical Profession and the NHS*, London: The Stationery Office.

Department of Health (2002a), *National Service Frameworks: a Practical Aid to Implementation in Primary Care*, London: The Stationery Office.

Department of Health (2002b), Investing in Primary Care, http://www.doh.gov.uk/pricare/investment/index.htm, updated 3 December 2002.

Department of Health (2003), Primary Care Act, Personal Medical Services (PMS), http://www.doh.gov.uk/pricare/pca.htm, updated 22 January 2003.

Dionne, G. and A-P. Contandriopoulos (1985), 'Doctors and Their Workshops: a Review Article', *Journal of Health Economics* 4, 21-33.

Dixon, J., P. Holland and N. Mays (1998), 'Developing Primary Care: Gatekeeping, Commissioning and Managed Care', *BMJ* 317, 125-128.

Dixon, P., H. Gravelle, H., R. Carr-Hill and J. Posnett (1997), *Patient Movements and Patient Choice, Report for National Health Service Executive*, York: York Health Economics Consortium.

Dranove, D., D. Kessler, M. McClellan and M. Satterthwaite (2002), Is More Information Better? The Effects of 'Report Cards' on Health Care Providers, NBER Working Paper 8697.

Dranove, D. and M. Satterthwaite (1992), 'Monopolistic Competition when Price and Quality are Imperfectly Observable', *RAND Journal of Economics* 23, 518-534.

Dranove, D. and M. Satterthwaite (2000), 'The Industrial Organization of Health Care Markets', in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics (Vol. 1)*, Amsterdam: North-Holland.

Dranove, D. and W.D. White (1987), 'Agency and the organization of health care delivery', *Inquiry* 24, 405-415.

Dranove, D. and W.D. White (1994), 'Recent Theory and Evidence on Competition in Hospital Markets', *Journal of Economics and Management Strategy* 3, 169-209.

Dusheiko, M., H. Gravelle, R. Jacobs, M. Kuhn and P. Smith (2001), Budgetary Devolution in the New NHS: Some Theoretical Economic Perspectives, paper presented at the Health Economists Study Group meeting, January 2001, Oxford.

Dusheiko, M., H. Gravelle, R. Jacobs and P. Smith (2002), The Effect of Budgets on Doctor Behaviour: Evidence from a Natural Experiment, paper presented at the Health Economists Study Group meeting, January 2003, Leeds.

Eastham, J. and B. Ferguson (2003), The economics of network forms and evaluation of clinical networks in the delivery of UK health care, paper presented at Health Economists Study Group meeting, January 2003, Leeds.

Edwards, N., M.J. Kornacki and J. Silversin (2002), 'Unhappy Doctors: What are the Causes and What Can Be Done?', *BMJ* 324, 835-838.

Ellis, R.P. (1998), 'Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins', *Journal of Health Economics* 17, 537-555.

Ellis, R.P. and T.G. McGuire (1986), 'Provider Behaviour under Prospective Reimbursement', *Journal of Health Economics* 5, 129-151.

Ellis, R.P. and T.G. McGuire (1993), 'Supply-Side and Demand-Side Cost Sharing in Health Care', *Journal of Economic Perspectives* 7, 135-151.

Elster, J. (1989), 'Social Norms and Economic Theory', *Journal of Economic Perspectives* 4, 99-117.

Emons, W. (1997), 'Credence Goods and Fraudulent Experts', *RAND Journal of Economics* 28, 107-119.

Encinosa III, W.E., M. Gaynor and J.B. Rebitzer (1997), The Sociology of Groups and the Economics of Incentives: Theory and Evidence on Compensation Systems, NBER Working Paper 5953.

European Observatory on Health Care Systems (2000), *Health Care Systems in Transition: Germany*, Copenhagen: World Health Organization.

Fleming, D. (1992), 'The Interface between General Practice and Secondary Care in Europe and North America' in: M. Roland and A. Coulter (eds.), *Hospital Referrals*, Oxford: Oxford University Press.

Frey, B.S. (1992), 'Tertium Datur: Pricing, Regulation and Intrinsic Motivation', *Kyklos* 45, 161-184.

Frey, B.S. (1997), 'On the Relationship between Intrinsic and Extrinsic Work Motivation', *International Journal of Industrial Organisation* 15, 427-439.

Garber, A.M., M.C. Weinstein, G.W. Torrance and M.S. Kamlet (1996), 'Theoretical Foundations of Cost-Effectiveness Analysis', in M.R. Gold, J.E. Siegel, L.B. Russell and M.C. Weinstein (eds.), *Cost-Effectiveness in Health and Medicine*, New York and Oxford: Oxford University Press.

Gaynor, M. (1994), 'Issues in the Industrial Organization of the Market for Physician Services', *Journal of Economics and Management Strategy* 3, 211-255.

Gaynor, M. and P. Gertler (1995), 'Moral Hazard and Risk Spreading in Partnerships', *RAND Journal of Economics* 26, 591-613.

Gaynor, M. and M.V. Pauly (1990), 'Compensation and Productive Efficiency in Partnerships: Evidence from Medical Group Practice', *Journal of Political Economy* 98, 544-573.

Getzen, T. (1984), 'A "Brand" Name Theory of Medical Group Practice', *Journal of Industrial Economics* 33, 199-215.

Giuffrida, A. and H. Gravelle (1998), 'Paying Patients to Comply: an Economic Analysis', *Health Economics* 7, 569-579.

Giuffrida, A. and H. Gravelle (2001), 'Inducing or Restraining Demand: the Market for Night Visits in Primary Care', *Journal of Health Economics* 20, 755-779.

Giuffrida, A., H. Gravelle and M. Roland (1999), 'Measuring Quality of Care with Routine Data: Avoiding Confusion between Performance Indicators and Health Outcomes', *BMJ* 319, 94-98.

Giuffrida, A., H. Gravelle and M. Roland (2000), 'Performance Indicators for Primary Care: The Confounding Problem', in P.C. Smith (ed.), *Reforming Markets in Health Care. An Economic Perspective*, Buckingham and Philadelphia: Open University Press.

Glazer, J. and A. Shmueli (1995), 'The Physician's Behaviour and Equity under a Fundholding Contract', *European Economic Review* 39, 781-785.

Glied, S. (2000), 'Managed Care', in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics (Vol. 1)*, Amsterdam: North-Holland.

Godber, E., R. Robinson and A. Steiner (1997), 'Economic Evaluation and the Shifting Balance towards Primary care: Definitions, Evidence and Methodological Issues', *Health Economics* 6, 275-294.

Goddard, M. and R. Mannion (1998), 'From Competition to Co-operation. New Economic Relationships in the National Health Service', *Health Economics* 7, 105-119.

Goddard, M., R. Mannion and P.C Smith (2000), 'The Performance Framework: Taking Account of Economic Behaviour', in P.C. Smith (ed.), *Reforming Markets in Health Care. An Economic Perspective*, Buckingham and Philadelphia: Open University Press.

Gold, M.R., D.L Patrick, G.W. Torrance, D.G. Fryback, D.C. Hadorn, M.S. Hadorn, M.S. Kamlet, N. Daniels and M.C. Weinstein (1996), 'Identifying and Valuing Outcomes', in M.R. Gold, J.E. Siegel, L.B. Russell and M.C. Weinstein (eds.), *Cost-Effectiveness in Health and Medicine*, New York and Oxford: Oxford University Press.

Goodwin, N. (1998), 'GP Fundholding', in J. Le Grand, N. Mays and J-A. Mulligan (eds.), *Learning from the NHS Internal Market. A Review of the Evidence*, London: The King's Fund.

Gosden, T., F. Forland, I.S. Kristiansen, M. Sutton, B. Leese, A. Giuffrida, M. Sergison and L. Pedersen (2001), 'Impact of Payment Method on Behaviour of Primary Care Physicians: a Systematic Review', *Journal of Health Services Research and Policy* 6, 44-55.

Gravelle, H. (1999), 'Capitation Contracts: Access and Quality', *Journal of Health Economics* 18, 315-340.

Gravelle, H., C. Bojke and B. Sibbald (2002a), 'Compensating Differentials for General Practitioners?' Paper presented at CEPR workshop on Health Economics and Public Policy 2002, Bergen.

Gravelle, H., M. Dusheiko and M. Sutton (2002b), 'The Demand for Elective Surgery in a Public System: Time and Money Prices in the UK National Health Service', *Journal of Health Economics* 21, 423-449.

Gravelle, H. and G. Masiero (2000), 'Quality Incentives in a Regulated Market with Imperfect Information and Switching Costs: Capitation in General Practice', *Journal of Health Economics* 19, 1067-1088.

Gravelle, H. and M. Sutton (2001), 'Inequalities in the Geographical Distribution of GPs in England and Wales 1974-1995', *Journal of Health Services Research and Policy* 6, 6-13.

Greenhalgh, T. and J. Eversley (1999), *Quality in General Practice*, London: The King's Fund.

Ham, C. and K.G.M.M. Alberti (2002), 'The Medical Profession, the Public, and the Government', *BMJ* 324, 838-842.

Hibbard, J. and J. Jewett (1997), 'Will Quality Report Cards Help Consumers?' *Health Affairs* 16, 218-228.

Hippisley-Cox, J., M. Pringle, C. Coupland, V. Hammersley and A. Wilson (2001), 'Do Single Handed Practices Offer Poorer Care? Cross Sectional Survey of Processes and Outcomes', *BMJ* 323, 320-323.

Hirshleifer, J. and J.G. Riley (1992), *The Analytics of Uncertainty and Information*, Cambridge: Cambridge University Press.

Huntington, J., S. Gillam and R. Rosen (2001), 'Clinical Governance in Primary Care: Organisational Development for Clinical Governance', *BMJ* 321, 679-682.

Iversen, T. and H. Lurås (2000), 'The Effect of Capitation on GPs' Referral Decisions', *Health Economics* 9, 199-210.

Jankowski, R.F. (2001), 'Implementing National Guidelines at Local Level: Changes in Clinicians' Behaviour in Primary Care Need to be Reflected in Secondary Care', *BMJ* 322, 1258-1259.

Juarez Garcia, A., R. Atun and J. Lord (2002), Economic Incentives and GPs as Purchasers of Elective Surgery: an Empirical Appraisal, paper presented at the Health Economists Study Group meeting, January 2003, Leeds.

Kihlstrom, R. and M. Riordan (1984), 'Advertising as a Signal', *Journal of Political Economy* 92, 427-450.

Konrad K. (2002), Competition for Prescription Drugs, paper presented at CEPR workshop on Health Economics and Public Policy 2002, Bergen.

Krasnik, A., P.P. Groenewegen, P.A. Pedersen, P. von Scholten, G. Mooney, A. Gottschau, H.A. Flierman and M.T. Damsgaard (1990), 'Changing Remuneration Systems: Effects on Activity in General Practice', *BMJ* 300, 1698-1701.

Kreps, D.M. (1990), 'Corporate Culture and Economic Theory', in J. Alt and K. Shepsle (eds.), *Perspectives on Positive Political Economy*, Cambridge: Cambridge University Press.

Kuhn (2001), Delegation and Regulation in the Provision of Health Care, poster presented at International Health Economics Association conference 2001, York; paper available from the author.

Le Grand, J., N. Mays and J-A. Mulligan (eds.) (1998), *Learning from the NHS Internal Market. A Review of the Evidence*, London: The King's Fund.

Lerner, C. and K. Claxton (1994), Modelling the Behaviour of General Practitioners. A Theoretical Foundation for Studies of Fundholding, Centre for Health Economics, University of York, Discussion Paper 116.

Lu, M., C-T.A. Ma, and L. Yuan (2000), 'Risk Selection and Matching in Performance-based Contracting', mimeo.

Lundin, D. (2000), 'Moral Hazard in Physician Prescription Behavior', *Journal of Health Economics* 19, 639-662.

Ma, C-T.A. (1994), 'Health Care Payment Systems: Cost and Quality Incentives', Journal of Economics and Management Strategy 3, 93-112.

McColl, A., P. Roderick, E. Wilkinson and J. Gabbay (2000), 'Clinical Governance in Primary Care Groups: the Feasibility of Deriving Evidence-based Performance Indicators', *Quality in Health Care* 9, 90-97.

McGuire, T.P. (2000), 'Physician Agency', in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics (Vol. 1)*, Amsterdam: North-Holland.

Mennemeyer, S.T., M.A. Morrisey and L.Z. Howard (1997), 'Death and Reputation: How Consumers Acted upon HCFA Mortality Information', *Inquiry* 34, 117-128.

Milgrom, P. and J. Roberts (1990), 'The Economics of Modern Manufacturing: Technology, Strategy, and Organization', *American Economic Review* 80, 511-528.

Milgrom, P. and J. Roberts (1992), *Economics, Organization and Management*, Upper Saddle River, NJ: Prentice Hall.

Naylor, R. (1990), 'A Social Custom Model of Collective Action', *European Journal of Political Economy* 6, 201-216.

Oxley, H. and M. MacFarlan (1994), *Health Care Reform. Controlling Spending and Increasing Efficiency*, Paris: OECD.

Pauly, M.V. (1980), *Doctors and Their Workshops: Economic Models and Physician Behaviour*, Chicago: University of Chicago Press.

Phelps, C.E. (2000), 'Information Diffusion and Best Practice Adoption' in A.J. Culyer and J.P. Newhouse (eds.), Handbook *of Health Economics (Vol. 1)*, Amsterdam: North-Holland.

Pringle, M. (1998), *Primary Care: Core Values*, London: BMJ Group.

Propper, C. (1993), 'Quasi-Markets and Regulation', in J. LeGrand and W. Bartlett (eds.), *Quasi-Markets and Social Policy*, Basingstoke and London: Macmillan.

Ratto, M., with S. Burgess, B. Croxson, I. Jewitt and C. Propper (2001), Team-based Incentives in the NHS. An Economic Approach, University of Bristol, CMPO Discussion Paper 01/37.

Rochaix, L. (1989), 'Information Asymmetry and Search in the Market for Physicians' Services', *Journal of Health Economics* 8, 53-84.

Rochaix, L. (1998), 'Performance-tied Payment Systems for Physicians', in R.B. Saltman, J. Figueras and C. Sakellarides (eds.), *Critical Challenges for Health Care Reform in Europe*, Buckingham and Philadelphia: Open University Press.

Roland, M. (1992), 'Measuring Appropriateness of Hospital Referrals', in M. Roland and A. Coulter (eds.), *Hospital Referrals*, Oxford: Oxford University Press.

Rosen, R. (2000), 'Clinical Governance in Primary Care: Improving Quality in the Changing World of Primary Care', *BMJ* 221, 551-554

Sachverständigenrat (2001), Appropriateness and Efficiency, English summary of volumes 1 and 2 of the report of the Advisory Council for the Concerted Action in Health Care, Bonn: Sachverständigenrat für die Konzertierte Aktion im Gesundheitswesen. Available online at http://www.svr-gesundheit.de/gutacht/gutalt/gutaltle.htm.

Saltman, R.B. and J. Figueras (1997), *European Health Care Reform. Analysis of Current Strategies*, Copenhagen: World Health Organization.

Scarpa, C. (1999), 'The Theory of Quality Regulation and Self-regulation: towards an Application to Financial Markets', in B. Bertolotti and G. Fiorentini (eds.), *Organized Interests and Self-regulation. An Economic Approach*, Oxford: Oxford University Press.

Scott, A. (1996), 'Primary or Secondary Care? What Can Economics Contribute to Evaluation at the Interface?' *Journal of Public Health Medicine* 18, 19-26.

Scott, A. (2000), 'Economics of General Practice', in A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics (Vol. 1)*, Amsterdam: North-Holland.

Scott, A. (2001), 'Eliciting GPs' Preferences for Pecuniary and Non-pecuniary Job Characteristics', *Journal of Health Economics* 20, 329-347.

Seddon, M.E., M.N. Marshall, S.M. Campbell and M.O. Roland (2001), 'Systematic Review of Studies on Clinical Care in General

Practice in the United Kingdom, Australia and New Zealand', *Quality in Health Care* 10, 152-158.

Shapiro, C. (1986), 'Investment, Moral Hazard, and Occupational Licensing', *Review of Economic Studies* 53, 843-862.

Smith, P.C. (1995), 'On the Unintended Consequences of Publishing Performance Data in the Public Sector', *International Journal of Public Administration* 18, 277-310.

Spence, M. (1973), 'Job Market Signalling', *Quarterly Journal of Economics* 87, 355-374.

Tannenbaum, S.J. (1993), 'What Physicians Know', *New England Journal of Medicine* 329, 1268-1271.

Thomas, P., F. Griffiths, J. Kai and A. O'Dwyer (2001), 'Networks for Research in Primary Care', *BMJ* 322, 588-590.

Tirole, J. (1986), 'Hierarchies and Bureaucracies: on the Role of Collusion in Organizations', *Journal of Law, Economics, and Organization* 2, 181-214.

Tirole, J. (1988), *The Theory of Industrial Organization*, Cambridge, Mass. and London, UK: The MIT Press.

Tirole J. (1994), 'The Internal Organization of Government', *Oxford Economic Papers* 46, 1-29.

Tirole, J. (1996), 'A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality)', *Review of Economic Studies* 63, 1-22.

Vick, S. and A. Scott (1998), 'Agency in Health Care. Examining Patients' Preferences for Attributes of the Doctor-Patient-Relationship', *Journal of Health Economics* 17, 587-605.

Wilkin, D. (1992), 'Patterns of Referral: Explaining Variation', in M. Roland and A. Coulter (eds.), *Hospital Referrals*, Oxford: Oxford University Press.

Wilkin, D. and C. Dornan (1990), *General Practitioner Referrals to Hospital: a Review of Research and its Implications for Policy and Practice*, Centre for Primary Care Research, University of Manchester.

Wilkin, D. and T. Smith (1987), 'Variation in General Practitioners' Referral Rates to Consultants', Journal *of the Royal College of General Practitioners* 37, 350-353.

Wilkinson, E., A. McColl, M. Exworthy, P. Roderick, H. Smith, M. Moore and J. Gabbay (2000), 'Reactions to the Use of Evidence-based Performance Indicators in Primary Care: a Qualitative Study' *Quality in Health* Care 9, 166-174.

Wolinsky, A. (1993), 'Competition in a Market for Informed Experts' Services' *RAND Journal of Economics* 24, 380-398.

Zweifel, P. and F. Breyer (1997), *Health Economics*, New York and Oxford: Oxford University Press.

Zweifel, P. and R. Eichenberger (1992), 'The Political Economy of Corporatism in Medicine: Self-regulation or Cartel Management?' *Journal of Regulatory Economics* 4, 89-108.