UNIVERSITÀ DEGLI STUDI DELLA CALABRIA
DOTTORATO «IMPRESA STATO E MERCATO»
XVII CICLO

TESI DI DOTTORATO
SETTORE SCIENTIFICO-DISCIPLINARE SECS/P02

QUALITY AND EFFICIENCY IN PRIMARY CARE THEORY AND EMPIRICAL EVIDENCE

SUPERVISORE
CH.MO PROF. FERDINANDO GROSSI

DOTTORANDA
GIULIANA DE LUCA

COORDINATORE
CH.MO PROF. GUERINO D'IGNAZIO

ANNO ACCADEMICO 2005 – 2006

INDEX

CHAPTER I

QUALITY INCENTIVES IN PRIMARY CARE - THEORY

CHAPTER II

QUALITY INCENTIVES IN PRIMARY CARE - EMPIRICAL EVIDENCE

CHAPTER III

DOES BETTER CLINICAL QUALITY IN PRIMARY CARE REDUCE ADMISSIONS FOR AMBULATORY CARE SENSITIVE CONDITIONS?

FIGURES INDEX

TABLES INDEX

ABSTRACT

Both researchers and policy-makers increasingly recognise to primary care a central role in the process of resource allocation as well as in the provision of high quality health care. Specifically, a great deal of the existing debate resolves around the role that the primary care physicians play in the production of care and as agents of both individual patients and the third payer. This thesis provides an overview of this debate and drawing on physician incentive literature tries to understand to which extent the trade-off between cost-containment and quality which affects the health care production may be mitigated.

The introductory section aims to outline the main features associated with the existing primary care systems. Analysing the institutional setting is as important as assessing the role of the physician in the production of health care. We start offering a definition of primary care and reviewing the organization and provision of services in general practice across European countries. A wide range of organisational structures for primary care exist in Europe. Overall, due to a lack of comparable data, the picture that may be sketched of primary care in the EU is partial. Within Europe, general practices differ in size, in payment method, in absence/presence of gatekeeper role, etc. The nature of the relationship between payers and doctors varies considerably across health systems. Three broad types of relationships can be identified: a *reimbursement approach*, with no formal relationship between the two parties; a *contractual approach*, with a contract between the two parties concerning the method and level of payments; an *integrated approach*, where medical practitioners are employed by insurers. Most national health systems pay general practitioners on the basis of a capitation system and patients are registered in a list. Others are paid by fee-for-services and the rest by salary. Actually, several countries have introduced a mixed payment and/or a pay-for-performance mechanism. In the following we will discuss how the organization of a health care system, and more generally, the institutional setting are an important determinant of the way in which primary care services are provided.

The remainder of the thesis is organised as follows. Chapters I and II deals with quality incentives from a theoretical and empirical perspective, respectively. Chapter III consists of an empirical evidence we carried out to assess the validity of using hospitalisations for so-called 'ambulatory care sensitive conditions' as an indicator of quality in primary care.

Chapter I provides an overview of quality incentives from a theoretical perspective. We start giving an operational definition for quality by applying Kuhn's distinction between 'structure', 'process' and 'outcome' aspects of quality in primary care adapted from Donabedian's framework. Then, we discuss advantages and disadvantages of two indicators researchers have developed and applied to assess quality on primary care: indicators derived from clinical audit and those based on ambulatory care sensitive conditions. The analysis is then carried out by considering how the institutional setting influences the physician decision process on investment and effort – which translate into quality - through the incentive structure. The degree of competition as well as regulation are important sources of quality incentives. These two sources are particularly stressed in the Health Economics literature. In addition, a third source of incentive is recognised especially by behavioural economics and sociology literature. This is given by values and the social norms. Most of the models on physician behaviour reviewed in this thesis are based on a utility-maximising framework which considers income as the main component that drives the physician decision process, but other aspects matter for a physician, i.e. the status as well as the patient's benefit and altruism. The remainder of the chapter draws on the contributions of several researchers to illustrate a framework to describe the physician's behaviour and how he influences both quantity and quality of services. This framework is developed stepwise, starting with models of complete information where non-retradability of medical services and non-contractible input (quality or effort) lead profit-maximizing physicians to influence the quantity demanded by patients. Subsequently, the third way of influencing patient demand is discussed: persuasion. Persuasion can be used by physicians in order to increase demand and thus maximise income when the physician does not act as a

perfect agent. In that light, explicit attention is paid to physician-induced demand (PID), one of the most discussed and examined issues in health economics. Attention is then paid to the alternative objective of a target income. The target income hypothesis is often used to explain PID (i.e. demand may be induced in order to ensure an acceptable income-level for the physician).

Subsequently, we move on to describe how the above mentioned sources of incentives work and how they can be modulated to induce the physician to provide efficient services and/or quality and, more generally, to improve the agency relationship. First, we discuss competition as a source of incentive. We examine two models, namely price and quality competition (model 1) and quality competition (model 2). In general, physicians may compete on price and/or on quality but, in practice, since in several countries the health care markets are highly regulated, the non-price competition is more likely to rise. Competition stimulates income related incentives if a physician's remuneration increases with the demand for his services and if this demand is reactive to quality. The demand elasticity to quality may be increased by reducing the switching cost from one physician to another, increasing the density of doctors, raising the outside options of care and reducing asymmetric information between physician and patient. Mechanisms to reduce asymmetric information are also discussed.

Then, we examine regulation as a source of incentives and we distinguish between soft and hard regulation. The former refers mainly to the group of the traditional payment mechanisms while the latter includes the so-called pay-for-performance mechanism. The debate on soft regulation, in particular, is mainly based upon the evaluation of different payment systems, such as prospective and retrospective payment systems or a mix of both, designed by the payer in order to induce the provider to supply a desired behaviour. We provide an overview of this debate and then we argue in favour of a mixed reimbursement as a mechanism able to mitigate the conflict between quality and efficiency. To this purpose, we provide a framework to compare purely retrospective and purely prospective payment systems, with respect to both micro-efficiency and macro-efficiency dimensions. The analysis is then carried out by focusing on the advantages of the mixed payment system over the two types mentioned above. Moreover, we

discuss which issues generally need to be taken into account when drawing up a doctor's reimbursement. The analysis is intended to be normative. In other words, it considers what is appropriate for a government purchaser to do when he is concerned with social welfare. Under existing payment systems, physicians usually receive the same payment regardless of the quality of care provided to patients. In other terms, quality incentives are generally implicit in the payment system we discussed above. To alter this equation, quality-related performance pay has been introduced in a number of health plans. The regulator establishes an explicit link between quality and the physician's income. The major problem recognised by the general literature on incentive is that payment for performance leads to difficulties in presence of multitasking. Moreover, accurate measure for physician actions that promote quality are extremely difficult to quantify. Using appropriate measures for quality is essential in designing a pay-for-performance system. To this purpose, the third chapter tries to face this issue.

Finally, we discuss the third source of incentives: values and social norms. We argue how a strong system of ethics and intrinsic motivation may weaken the incentives implicit in several payment systems. There is also risk of crowding out if the external incentives are seen as substitutes for intrinsic motivation.

Chapter II deals with the empirical evidence on physician quality incentives. Specifically, we start by highlighting the methodological issues to be addressed when critically appraising the evidence and, subsequently, we address the question on what the empirical evidence tell us about provider's responses to financial incentives. The bulk of reviewed papers concerns with the impact of payment systems on behaviour of physicians. Only a few high-quality studies are illustrated despite the great number of papers written on this topic. Several empirical studies are based on natural experiments and are therefore opportunistic. Only a few studies have been able to control accurately for differences between physicians that may influence their behaviour - such as patient characteristics, case mix, physician characteristics, geographical and social factors, etc.. Then, we present the evidence derived from several studies carried out in different contexts. The principal lesson we draw from this studies is that incentive design matters.

For example, Krasnik et al. (1990) provides strong evidence that changing from capitation payment to fee-for-service payment presents income-generating opportunities of which doctors take advantage through increasing service intensity and workload. Giuffrida and Gravelle (2001) present evidence that the number of doctor visits responds positively to an increase in fees. There is evidence of a reduction in the use of deputising services when differential fees were introduced as part of the 1990 doctor contract in UK. Rochaix (1993) in analysing the impact of price-quantity regulation on the overall level of activity and the mix of services provided by general practitioners in Québec finds that a fee freeze is associated with increased activity and a drift to more complex procedures. Yet, the empirical evidence specifically on quality in primary care is limited and often appears inconclusive despite the number of studies on the impact of payment systems on physician behaviour. Also the empirical bases of pay-for- performance in health care are rather weak. There are only a few studies demonstrating that pay-for-performance leads to improved quality of care.

Chapter III illustrates an empirical work dealing with the measuring of quality in primary care. The title is '*Does Better Clinical Quality in Primary Care reduce Admissions for Ambulatory Care Sensitive Conditions?* '. Measuring quality has been characterized by several empirical studies in the literature dealing with performance in the primary care system. Mainly, two different sets of indicators have been applied: the most common is derived from hospital statistics (admission rates for ambulatory care sensitive conditions -ACSCs) whereas the other is from medical record based on clinical audit (patient/practice quality scores). This study tries to assess the validity of hospitalisations for ambulatory care sensitive conditions as an indicator of clinical (or technical) quality in primary care. The setting is a longitudinal study of a cohort of 60 family practices in England over two years (1998 and 2003). We found that clinical quality of care explains a small proportion of the variation in admission rates, and its effect is not always in the expected direction. Caution is required when using ambulatory care sensitive admission rates as an indicator of clinical quality in primary care.

SOMMARIO

Background Il settore della medicina di base (o della cura primaria o *primary care*) è parte integrante di ogni sistema sanitario e, più in generale, dell'economia di un paese. Il medico di base costituisce il primo livello di contatto per gli individui che esprimono un bisogno in termini di miglioramento del proprio stato di salute e, in diversi paesi europei, ha il compito di commissionare l'utilizzo della cura secondaria. Il ruolo cruciale svolto dal medico nell'assicurare appropriati livelli di qualità nella cura della salute di un paziente, e più in generale, dell'intera popolazione è ben riconosciuto tanto dalla produzione scientifica internazionale quanto dai decisori politici.

Obiettivi Il principale obiettivo del presente lavoro di tesi è volto ad analizzare i principali meccanismi di incentivazione della qualità nell'erogazione delle prestazioni di medicina di base, soffermandosi su quelli economici e sulle implicazioni in termini di *policy*. Questi argomenti sono illustrati attraverso esempi e formalizzazioni (Capitolo I) e accompagnati dall'evidenza empirica prodotta da ricercatori in diversi contesti internazionali (Capitolo II). All'analisi della letteratura segue un contributo di natura empirica sulla relazione tra indicatori di *outcome* e indicatori di processo tradizionalmente utilizzati per la misurazione della qualità in medicina generale (Capitolo III).

Struttura Si parte da una definizione di qualità nel settore della medicina di base e, quindi, si prosegue con la distinzione tra incentivi monetari (interni ed esterni) e non monetari alla qualità. I primi derivano dall'ipotesi di massimizzazione del reddito da parte del medico mentre i secondi derivano dall'ipotesi che il medico includa nella propria funzione di utilità altri aspetti quali lo status, le motivazioni intrinseche e l'altruismo. Vengono analizzate due categorie di incentivi monetari. La prima attiene agli incentivi c.d. interni derivanti dal mercato (in particolare, ruolo dell'elasticità della domanda delle prestazioni mediche e della competizione nell'erogazione della qualità sotto l'ipotesi che il medico riceva una remunerazione fissa). In questo contesto, si discute il problema dell'informazione nascosta relativa all'aspetto della qualità collegato alla struttura e quello dell'azione nascosta relativo al processo. Si analizzano le soluzioni al problema dell'asimmetria informativa (di *search*,

segnalazione, reputazione, misure di regolamentazione, accreditamento, certificazione e uso di indicatori di performance). Si esaminano quindi gli incentivi monetari esterni derivanti dalla regolamentazione e distinguendo anche qui tra regolamentazione indiretta (definizione dei meccanismi di rimborso – *capitation, salary, fee-for-services*) e diretta (sistemi di remunerazione basati sulla *performance*). Tra gli incentivi non monetari vengono discussi quelli derivanti dall'altruismo, dalle motivazioni intrinseche, dall'interazione sociale. La tesi si conclude con un lavoro empirico sull'opportunità di misurare la qualità della medicina di base utilizzando indicatori di *outcome* piuttosto che di processo.

Letteratura In una prospettiva positiva, gli economisti spiegano il livello di qualità scelto dal medico di base come una risposta agli incentivi derivanti dall'assetto istituzionale. Questo, a sua volta, è definito dai sistemi di remunerazione e dagli standard di performance fissati dal decisore politico. In una prospettiva normativa, invece, gli economisti si domandano quale combinazione di investimenti, di sforzo e di *inputs* ottimizzi la qualità, dato il vincolo delle risorse disponibili. Soggetto ai vincoli derivanti dal mondo reale (per esempio relativi all'informazione, alla stipulazione del contratto o alla definizione delle politiche) l'economista e, successivamente, il decisore politico si domandano quindi quale assetto istituzionale promuovere allo scopo di implementare una soluzione che sia quanto più possibile vicina all'ottimo desiderato (Kuhn, 2003).

Nel tentativo di definire un assetto istituzionale appropriato, il decisore politico mira al perseguimento della micro-efficienza (massimizzare il guadagno atteso in termini di salute dato un set di risorse), della macro-efficienza (attraverso la ottimizzazione del livello delle risorse allocate alla cura primaria) e dell'equità. Conflitti tra questi obiettivi generalmente costringono il decisore politico a dover scegliere tra l'erogazione di servizi di alta qualità e il contenimento dei costi per un verso, e tra equità ed efficienza per l'altro. Dunque unitamente alla definizione del livello ottimale di qualità per un dato set di risorse destinato alla cura primaria, una questione diversa ma altrettanto rilevante è la scelta sul livello di risorse da destinarsi alla cura primaria stessa (macro-efficienza). Il criterio di efficienza a livello allocativo richiede che il decisore scelga un set di risorse tale che il beneficio marginale a livello micro bilanci i benefici derivanti dall'uso alternativo

delle risorse, cioè quelle impiegate nella cura secondaria, in altre aree della spesa pubblica o in investimenti privati o nel consumo. In definitiva, il decisore politico è interessato ad incentivare la qualità nell'ambito del settore della medicina di base, compatibilmente al vincolo di spesa. Dato questo vincolo il medico, a sua volta, realizzerà la sua scelta in termini di investimenti, sforzo e impiego di altri input derivanti, ad esempio, dal settore secondario e quello farmaceutico. Più in generale, l'assetto istituzionale - definito dalle condizioni di mercato, dai sistemi di remunerazione, dal rapporto con gli altri livelli di cura, ecc - insieme alle spinte motivazionali - che possono dipendere dal raggiungimento di un livello di reddito o status ma anche dalla realizzazione di valori intrinseci – guideranno il medico nel suo processo decisionale.

La necessità di predisporre delle misure per incentivare la qualità da parte del regolatore dipende principalmente dal fatto che il settore della medicina di base non è un settore perfettamente concorrenziale. In presenza di un soggetto massimizzatore del reddito, è importante far leva sul fatto che gli incentivi alla qualità sono tanto più forti quanto più elastica è la domanda di cura da parte del paziente rispetto alla qualità. L'elasticità della domanda rispetto alla qualità, a sua volta, è direttamente proporzionale alla densità dei medici[1] presenti in una determinata area e alla disponibilità dell'uso alternativo delle fonti di cura (per esempio, la telemedicina, l'auto-trattamento, la distribuzione dei farmaci per uso domestico o altri servizi medici alternativi)[2]. Inoltre essa diminuisce quanto più elevati sono i costi di *swithching*[3] e in presenza di problemi di asimmetria informativa sulla qualità di un servizio (Dranove e Satterthwaite 1992, 2000). In quest'ultimo caso un intervento del *policy maker* potrebbe contribuire al miglioramento della qualità attraverso, ad esempio, il rafforzamento delle opzioni di cura esterne alla medicina di base. Il National Health System (NHS) inglese, ad esempio, ha effettuato ingenti investimenti per migliorare l'informazione e

[1] Gravelle (1999) analizza la concorrenza per la qualità tra un certo numero di medici in un'area in cui i pazienti devono spostarsi per ricevere la cura. Sotto un regime di concorrenza, la qualità aumenta sia con la remunerazione che con il numero di medici presenti nell'area.
[2] La disponibilità di fonti alternative di cura riduce il potere di mercato del medico aumentando il grado di concorrenza. In questo modo si assicura una maggiore produzione di qualità.
[3] Elevati costi di trasporto per il paziente implicano che questi percepiscono i vari servizi offerti dai medici come poveri sostituti. La riluttanza a cambiare il proprio medico aumenta il suo potere di mercato e quindi indebolisce gli incentivi alla qualità.

l'accesso del paziente alla medicina di base (Department of Health, 2000)[4]. Purtroppo, la presenza di asimmetria informativa sulle abilità (*hidden information*) e sullo sforzo del medico (*hidden action*) e di costi di *swithching* restringono spesso le scelte del paziente riducendo la concorrenza (Hirshleifer and Riley, 1992). Questo molto probabilmente conduce a una sotto-produzione della qualità in generale, o alla produzione della qualità solo per quelle dimensioni che sono osservabili dai pazienti, per esempio le amenità (Chalkley e Malcomson, 1998, 2000; Dranove e Satterthwaite 1992, 2000). La presenza di asimmetria informativa tra medico e paziente può essere risolta attraverso vari meccanismi[5]. Il regolatore può scegliere di promuovere meccanismi di segnalazione o di reputazione incrementando la remunerazione percepita dal medico (modo indiretto) oppure implementando misure più dirette che includono l'*accreditamento*, ossia la definizione di standard minimi per poter svolgere l'attività di medico[6]; la *certificazione*, ossia la pubblicizzazione dell'investimento effettuato in capitale umano (Shapiro, 1986) e l'uso di indicatori di performance (Gravelle e Masiero, 2000). In generale, la risoluzione dell'asimmetria informativa risulta costosa da un punto di vista sociale. Questo costo deve essere considerato come un costo indiretto della qualità.

La natura del sistema di remunerazione o di rimborso ha conseguenze importanti nella produzione della qualità e il decisore politico può utilizzare uno schema di pagamento giudicato più appropriato per realizzare il livello desiderato di qualità. Dalla ricerca effettuata risultano alcune tendenze e caratteristiche comuni relativamente alla forma di remunerazione che i medici di base ricevono per le loro prestazioni e in base alla presenza/assenza della funzione di

[4] Alcuni esempi sono: NHS Direct via telefono, la costituzione di *information points* in luoghi pubblici e di *walk-in centers* costituiti da uno staff di infermieri piuttosto che di medici.

[5] I meccanismi discussi in letteratura sono principalmente i seguenti: di *search*, se il paziente è in grado di controllare gli aspetti rilevanti della qualità, ad esempio inferendo il livello di qualità dalle attrezzature disponibili o dalle certificazioni sulla professionalità del medico (Rochaix, 1989); di *signalling* dell'informazione nascosta, ad esempio attraverso l'acquisizione di certificati extra o l'acquisizione di una reputazione sulla qualità se la cura è considerata un '*experience good*' (ossia i pazienti sono in grado di determinare la qualità una volta che essi fanno uso del servizio); di *collective reputation* (reputazione collettiva) o credenziali professionali derivanti da esperti indipendenti se la cura è considerata un '*credence good*' (ossia i pazienti non sono in grado di giudicare la qualità). Se i meccanismi di reputazione falliscono l'unico modo per garantire la qualità è dato dai meccanismi di *screening* e *monitoring* del medico da parte di un controllore esperto. In questo caso l'informazione può essere usata direttamente per regolamentare il *provider* o essere trasmessa ai pazienti sotto forma di indicatori di performance.

[6] In questo senso l'accreditamento aiuta a mitigare il problema dell'informazione nascosta e se il costo addizionale nel produrre qualità diminuisce all'aumentare delle abilità possedute dal medico, l'accreditamento promuove anche la reputazione come garanzia della qualità e indirettamente può ridurre il problema dell'azzardo morale.

gatekeeping nell'ambito dei sistemi sanitari esistenti. I sistemi basati sull'assicurazione sociale (cosiddetto modello Bismark, presente in Germania, Austria, Belgio, Francia, Lussemburgo e Paesi Bassi) tendono ad essere caratterizzati da un sistema di remunerazione retrospettico e il medico non svolge la funzione di *gatekeeper* (cioè i pazienti hanno accesso diretto alla cura specialistica), mentre i sistemi sanitari nazionali basati sulla tassazione generale (cosiddetto modello Beveridge, presente in UK, Danimarca, Grecia, Irlanda, Italia, Norvegia, Spagna e Svezia) tendono ad essere caratterizzati da sistemi d remunerazione prospettici e da *gatekeeping*. Entrambe le forme di pagamento generano degli effetti distorsivi con rilevanti implicazioni per la qualità. Gli incentivi derivanti da queste due forme di pagamento sono rafforzati da quelli derivanti l'organizzazione dell'accesso alla cura specialistica. Sotto un regime di *gatekeeping*, che è spesso associato con la registrazione dei pazienti ad una lista, c'è un limitato interesse del paziente a rivolgersi a fornitori di cura alternativi quando non sono soddisfatti. In questa circostanza il grado di competizione tra i medici per assicurarsi un maggior numero di pazienti è ridotta. Al contrario, l'accesso diretto può indurre i medici a sovra-produrre i servizi domandati da pazienti assicurati e a non incoraggiarli a rivolgersi direttamente allo specialista. A seguito di alcune riforme alcuni paesi hanno preferito introdurre sistemi misti di pagamento (Rochaix, 1998) in sostituzione di sistemi prospettici e retrospettivi puri. Infatti, l'assegnazione al medico di base di un *budget fisso* incentiva la produzione di un mix di servizi efficiente ma conduce alla sotto-produzione degli stessi e della loro qualità se il medico ha l'obiettivo della massimizzazione del proprio reddito. Lo stesso vale quando la remunerazione è costituita da uno *stipendio fisso*. In questa circostanza si osserva un effetto distorsivo aggiuntivo nel mix dei servizi verso maggiori prescrizioni e rinvii dei pazienti alla cura specialistica. Un sistema basato su un rimborso fisso per ciascun paziente iscritto in una lista (*capitation*) offre corretti incentivi alla qualità, purché la domanda di tutti i pazienti sia reattiva alle diverse dimensioni della qualità. Infatti, se il medico massimizza una funzione di utilità che dipende principalmente dal reddito, questo schema di pagamento conduce a una sotto-produzione di quegli aspetti

della qualità non osservabili o alla discriminazione dei pazienti[7]. Il rimborso per ogni prestazione resa (*fee-for-services*) conduce invece ad una sovra-produzione dei servizi con ambigue implicazioni per la qualità. La qualità potrà risultare troppo elevata, cioè a livelli che non sono più costo-efficaci, o troppo bassa, dovuto alla distorsione nel mix di servizi scelti. Per esempio, sistemi basati su *fee-for-services* possono indurre i medici a trattare pazienti che dovrebbero essere rinviati ad uno specialista. Se i pazienti hanno accesso diretto alla cura specialistica, questo sistema di rimborso incentiverà i medici a specializzarsi allo scopo di differenziare i servizi offerti.

Di recente, diversi sistemi sanitari hanno introdotto in modo esplicito incentivi alla qualità (*performance pay*) negli schemi di remunerazione dei medici. Il vantaggio dei sistemi di remunerazione basati sulla performance sta nel fatto che essi contengono incentivi alla qualità anche se la domanda di prestazioni di cura da parte del paziente è anelastica rispetto alla qualità. Tuttavia, la predisposizione di schemi di performance implica una serie di problemi: 1) richiede che tutti gli aspetti della performance siano rimborsabili; 2) la creazione di incentivi di team; 3) il contenimento del rischio di performance a cui il medico è esposto; 4) la determinazione dell'appropriato grado di monitoraggio; 5) la determinazione dei *benchmarks* di performance; 6) la credibilità del regolatore nell'impegnarsi a mantenere gli standard nei contratti futuri (*ratchet effect*). Il nuovo contratto dei medici di base siglato nel 2004 in Inghilterra costituisce un ambizioso tentativo di incorporare gli incentivi alla qualità nella remunerazione del medico di base. Il meccanismo tradizionalmente utilizzato per favorire la qualità è rappresentato in molti sistemi sanitari dall'*exit*, cioè pazienti non soddisfatti cambiano il proprio medico di base riducendo la quota di reddito percepita dal medico essendo quest'ultimo basato sul numero degli iscritti nella lista dei pazienti. Nuovi

[7] Se alcuni pazienti manifestano una elasticità di domanda più bassa, ad esempio dovuta alla mancanza di informazione o perché hanno bisogno di un trattamento urgente o affrontano elevati costi di *search* i medici possono avere degli incentivi alla loro discriminazione e sotto-produrre la qualità in riferimento a questa tipologia di pazienti. Questo porta ad una perdita di efficienza e di equità (Allen e Gertler, 1991). Se, ad esempio, la remunerazione è calcolata come una media dei costi di trattamento per pazienti gravemente e non gravemente ammalati, il medico tenderà ad attrarre pazienti il cui costo di trattamento è minore della media (*cream-skimming*) e a discriminare quelli che generano costi di trattamento superiori alla media (*skimping*). Ellis (1998) mostra sotto quali condizioni pazienti la cui cura risulti particolarmente costosa siano scaricati (*dumped*). Una alternativa consiste nel rifiutarsi di iscrivere il paziente nella propria lista. Allo scopo di ridurre le diverse forme di discriminazione il regolatore dovrà definire degli schemi di pagamento specifici per ogni tipologia di paziente. Tuttavia questa *policy* non è praticabile a causa della scarsa informazione e dei costi amministrativi elevati. Il regolatore dovrebbe imporre standard minimi di qualità e promuovere campagne informative per quei pazienti la cui elasticità di domanda rispetto alla qualità dei servizi è particolarmente bassa.

contratti come quello che caratterizza la *NHS primary care* in Inghilterra invece affiancano a questo meccanismo incentivi diretti alla promozione della qualità (Roland, 2004).

Attraverso la pubblicazione di indicatori di performance, il regolatore migliorando l'informazione del paziente aumenta l'elasticità della domanda rispetto alla qualità e quindi la concorrenza. Se gli indicatori di performance scelti sono positivamente correlati con la qualità, questo obiettivo può essere raggiunto. Tuttavia, gli indicatori possono omettere la misurazione di alcune dimensioni della qualità giudicati importanti ma difficili da catturare. Se la produzione della qualità è più che la somma delle parti che vi concorrono, allora misurare la qualità attraverso un set di indicatori individuali risulta fuorviante. Inoltre, l'uso di indicatori può indurre il medico a porre l'attenzione solo sulle dimensioni della qualità che si richiede siano misurate, a manipolazioni, *gaming*, etc. (Smith, 1995).

Tradizionalmente, specie negli Stati Uniti, i tassi di ospedalizzazione per le cosiddette 'condizioni sensibili alla cura di base' (*Ambulatory Care Sensitive Conditions*) sono stati impiegati come indicatori per misurare l'accesso alla medicina di base (Billing et al, 1993, 1996; Bindman et al. 1995; Laditka et. al, 2005) e in Inghilterra come indicatore di *performance* del medico di base. Tuttavia, Giuffrida et. al (1999, 2000) dimostrano un limite intrinseco nell'uso di questo indicatore dovuto all'elevata variabilità nel tempo che costringe a calcolarli come *moving averages*. Inoltre, in qualità di indicatori di performance dovrebbero includere solo gli aspetti della cura che possono essere tenuti sotto il controllo del medico. Essi hanno dimostrato, tuttavia, che la valutazione della performance delle autorità sanitarie locali basata sui tassi di ospedalizzazione varia a seconda che questi siano utilizzati come tassi grezzi, o aggiustati per età e sesso; quindi per fattori di morbilità; per fattori socio-demografici; e per l'offerta di cura secondaria (ossia ospedaliera).

In diversi sistemi sanitari nazionali la *self-regulation* della qualità da parte del professionista è necessaria ma non sufficiente. Una possibile spiegazione è che il *free-riding* conduce ad una mancanza di incentivi per sostenere una reputazione collettiva (piuttosto che individuale). Alla luce di questo, di recente il governo

clinico (*clinical governance*) ha ricevuto una sostanziale attenzione dai *policy-makers* e dai ricercatori come meccanismo potenzialmente potente di controllo della qualità. È riconosciuto che la *clinical governance* comprenda un mix di regolazione esterna e *self-regulation*. Il governo clinico può essere inteso come un insieme di semplici regole formali e non finalizzato a promuovere comportamenti appropriati sotto contingenze non verificabili. A questo riguardo, esso facilita la creazione di reputazione (sia dei medici che del terzo pagante). Condivisione delle informazioni e apprendimento collettivo sono intesi come elementi chiave del governo clinico. Essi possono essere interpretati come una forma di regolamentazione partecipatoria, in cui i medici sono coinvolti nel determinare la propria performance. Mentre questo facilita il compito del regolatore, apre un canale per sprechi da un punto di vista della società (Kuhn, 2003). In Inghilterra, *Primary Care Trusts* implementano gli standard di performance introducendo un sistema di governo clinico. A questo riguardo, gli organi appartenenti ai *Primary Care Trusts* responsabili per il governo clinico assumono una funzione come supervisori in un'agenzia gerarchica.

L'altruismo e la piena considerazione del benessere del paziente, motivazioni intrinseche costituiscono un deterrente alla sotto-produzione della qualità anche se contengono un alto potenziale per l'uso inefficienze delle risorse. Tuttavia, l'interrelazione tra incentivi non finanziari derivanti dall'altruismo, da motivazione intrinseche e dall'interazione sociale e quelli finanziari non è stata oggetto di attenzione da parte dei ricercatori sia da una prospettiva teorica che empirica.

L'evidenza empirica supporta alcune predizioni teoriche sugli incentivi derivanti da schemi di remunerazioni differenti. Tuttavia, le implicazioni per la qualità della cura primaria sono state scarsamente esplorate. In generale, l'evidenza empirica sulla qualità nella medicina generale è limitata e spesso inconcludente (Krasnik et al, 1999; Iversen and Luras 2000; Giuffrida e Gravelle 2001; Juarez et al., 2002; Gravelle et al. 2002; Dusheiko et al. 2003. Il problema di trovare buone misure della qualità caratterizza tutti i lavori empirici (Kuhn, 2003).

L’evidenza contenuta nel Capitolo III ha cercato di contribuire a colmare questo gap. A questo scopo viene testata l’ipotesi che una migliore qualità di processo nel trattamento di tre condizioni c.d. sensibili alla cura erogata attraverso i servizi di medicina di base – diabete, asma, angina – si rifletta in un migliore esito per la salute di un paziente e dunque riduca i relativi tassi di ospedalizzazione, generando un risparmio di risorse impiegate nel settore di cura secondario. Disponendo di dati sulla qualità clinica dei medici di base per l'Inghilterra, il lavoro testa l'ipotesi che i tassi di ammissione ospedaliera aggiustati per fattori non controllabili dal medico (di morbilità, demografici, socioeconomici, geografici, di offerta di cura secondaria) siano inversamente associati con la qualità. Tuttavia, i risultati ottenuti non sono univoci. Per il diabete è stata stimata una relazione positiva. Per l'angina il segno dell'associazione è quello atteso, ossia l'analisi mostra un'associazione negativa tra tassi di ammissione ospedaliera e la qualità clinica. Non è stata trovata alcuna associazione tra i tassi di ammissione ospedaliera e gli score di qualità in relazione all’asma. Si raccomanda, dunque, una certa cautela nell’impiego dei tassi di ammissione ospedaliera per le suddette patologie come misura indiretta della qualità clinica in medicina di base.

INTRODUCTION

Because of the trade-off between efficiency and quality which affects the health care production, provider incentive literature has become a relevant topic in health economics. Over the last decades, both researchers and policy-makers have especially paid attention on the role of primary health care in the process of resource allocation and in the provision of high quality health care. In this thesis we review several models - both theoretical and empirical – which provide insightful thoughts about the issue of general practitioners behaviour in their role of assemblers of care and agent of both individual patients and the payer. Several definitions of both efficiency and quality have been developed and applied in the existing literature depending on the perspective of the researcher or the stakeholder. This thesis embrace Rochaix's definition of efficiency (1998) and Donabedian's definition of quality (1980). Rochaix' definition of efficiency covers two dimensions of efficiency, i.e. micro-efficiency and macro-efficiency, in order to evaluate the physician behaviour. Both of them involve the agency relationship. In particular, the physician is regarded as the agent of two principals. One principal is the patient, who is concerned to receive an appropriate treatment, and the other one is the payer (an insurer or a government agency), who is interested in the treatment that patients receive as well as keeping down the cost of providing treatments. Consequently, since the payer has a secondary aim which is cost-containment, a potential conflict of interest between the patient and the payer could arise. In the light of this tension, the concepts of micro-efficiency and macro-efficiency assume relevance. Micro-efficiency (in both allocative as well as productive meaning) does not involve actually any conflict between the payer and the patient, because the patient's preferences are taken into account. The type and the level of treatment are supposed to be very relevant and the use of factors be efficient for any given treatment. In contrast, macro-efficiency refers to the physician-third-party payer agency relationship and it essentially refers to cost-containment. On the macro-side, what matters is the efficient overall level and structure of health care given the scarcity of funds. In this case, the optimal level of care depends on the opportunity cost of spending the same funds on other sectors such as education, retirement, etc.

As for quality, a former definition describes it as the effect of care on the health of the individual and of the population (Rutstein et al., 1976). Even though Rutstein's definition is broad and partial (it refers only to the fact that improvement in quality should be reflected in better health), it captures both micro and macro aspects of quality. More recently, the concept of quality has been well specified and extended to include not only aspects related to the outcome but also to structure and process (Donabedian, 1980, Campbell et al., 2000). On the micro-side, what matters is measuring 'whether individuals can access the health structures and processes of care which they need and whether the care received is effective'(Campbell et al., 2000, p. 1614). Therefore, the two main components of quality of health care at the micro level are access and effectiveness. Access is mainly intended in terms of both geographic/physical access – the extent to which there exist barriers such as rurality, disability or old age to reaching health care facilities - and of availability - the extent to which the health care system provides facilities (structures) and services (process) to meet the needs of individuals. Effectiveness, instead, refers to both clinical - and interpersonal - care. The first deals with effective application of evidence-based or legitimate medicine while the second takes into account the fact that processes and outcomes depend on the specific characteristics of the patient. Co-ordination or integration of care for individual patients is also another aspect of effectiveness and is a further attribute of particular relevance to primary care (Starfield, 1994). By contrast, on the macro-side more relevant is 'the ability to access effective care on an efficient and equitable basis for the optimization of health benefit/well being for the whole population' (Campbell et al., 2000, p. 1617). Therefore, equity and efficiency are the additional components of quality for population.

Conflicts between the micro-efficiency an the macro-efficiency aims can arise, leading to trade-offs between high quality and cost-containment on one hand, and efficiency and equity on the other (Kuhn, 2003). Starting from this framework, several studies tried to develop and apply indicators with the aim to reflect different dimensions of quality in general practice from diverse perspectives, using different methods and sources of data. However, a discrepancy between the definition of the multidimensional aspects of care and the methods currently used

for collecting data to measure such aspects persists. For example, data contained in administrative datasets used for financial purposes cannot be applied to obtain a comprehensive quality assessment but for assessing those aspects of quality 'which are consistently and reliably recorded therein' (Campbell et al., 2000). Campbell et al. (2000) suggested using this framework to permit critical analysis of quality indicators and to show which aspects of quality they measure and which ones do not. Giuffrida et al. (1999) pointed out that a clear distinction need to be considered between health outcomes measures (such as admissions, deaths, disability) and performance indicators, the last ones being related with those aspects of care which can be affected by the staff and by organization of care more generally.

The remainder of this thesis is organized as follows. The next section aims to outline the main features associated with the existing primary care systems. This discussion allows us to understand how different institutional settings impact the organisation of primary care and through this the doctor behaviour. Chapter I analyses from a theoretical perspective the issue of quality incentives by highlighting the source from which they arise, namely competition, regulation and intrinsic motivation. Most of the models reviewed are based on a utility-maximising framework which considers income as the main component but this is not the only one. Other aspects, such as the patient's benefit and altruism, are also included. Chapter II deals with the empirical evidence produced on quality incentives. Several predictions made by the theory and tested empirically are reviewed. Specifically, the attention is paid on the quality of the study design. Thus, only a few studies are illustrated despite the great number of papers written on this topic. Chapter III illustrates an original empirical work carried out in the UK setting. This tries to assess how better technical quality in primary care translates in better health outcome, as measured by hospitalizations for ambulatory care sensitive conditions. Our results could help the policy maker for the valuation of the pay-for-performance payment system introduced through the new contract in 2004 in UK.

INSTITUTIONAL SETTING

This section provides a definition of primary care and reviews the organization and provision of services in general practice across European countries. Definitions of primary care are numerous and either descriptive or normative, depending on the purpose they serve. The normative approach has been closely linked with the World Health Organisation (WHO) Alma Ata Declaration in 1978 on primary health care:

'Primary health care is essential health care based on practical, scientifically sound and socially acceptable methods and technology made universally accessible to individuals and families in the community through their participation and at a cost that the community and country can afford to maintain at every stage of their development in the spirit of self-reliance and self-determination. It forms an integral part both of the country's health system, of which it is the central function and main focus, and of the overall social and economic development of the community. It is the first level of contact of individuals, the family and community with the national health system bringing health care as close as possible to where people live and work, and constitutes the first element of a continuing health care process' (WHO, 1978).

The reported definition includes several aspects of primary care. It focuses on solidarity and equitable access to care; on the protection and promotion of health rather than on treatment itself; it takes into account the influence of social, economic and environmental factors on health; and underline a broad intersectorial collaboration in dealing with community problems. Primary care is only a piece of the entire health care delivery system which is often represented as a pyramid. It occupies the interface between self-care and hospital-based secondary (general specialist) and tertiary (super specialist) care. Before primary health services are used, self care is widely practiced. Secondary care follows the primary care level while tertiary care is situated at the top of the pyramid (see Fig. 21 in the appendix). Informal care is an unspecified area below the pyramid.

Primary care is the response to unspecified and common health problems accounting for a vast majority of the population health needs. Problems that require more specialized medical expertise are dealt with in secondary care, in hospitals or the outpatient context, while rare and very complex cases are treated in tertiary care (Fry, 1992).

A number of health sector reforms during the 1990s have aimed to readjust the division of tasks between primary and secondary care. The role of primary care in managing the entrance to and exit from secondary care (or 'gatekeeping' role) has been enhanced. In addition, inpatient care is necessary for a diminishing fraction of those who enter the health care system and staying in hospitals have become shorter (White *et al.*, 1961; McKee and Healy, 2001). Most of the patient population can be treated in the primary care setting where up to 90 per cent of all health care activity can be performed (Hobbs, 1995). This figure highlights the need to reach an appropriate balance of resources between primary and secondary care, with adequate financial and human resources being directed towards the primary care sector, and in line with changes in the task division between primary and secondary care (Forrest and Starfield, 1996; Jepson, 2001).

Information on the allocation of resources in primary health care is not plentiful. Most studies have focused on the hospital sector rather than on primary care, perhaps in part because definitional problems related to primary care make comparison difficult. As a consequence there is no source of comparable data on, for example, the financial resources allocated to the primary care sector in Europe. Neither the OECD database nor the Health-for-All database do not provide these data. Some national databases provide information on resources devoted to primary care but differences in definitions, parameters, and data make comparison difficult (Lagasse *et al.*, 2001). Concerning human resources, data on primary care practitioners are also incomplete and often inadequate. Despite of the central role of human resources in the health sector, international attention to human resources for health has only recently emerged on the health policy agenda (Dubois *et al.*, 2005).

Existing data on expenditure in primary care suggest that less than a quarter of the health care budget generally goes to primary care in western Europe (Hobbs, 1995; Goicoechea, 1996; OHE, 2000). Overall, health care resources across Europe have increased in real terms over recent decades, reflected in increases in health expenditures. Indirect indicators, however, do not indicate that the proportion for primary care has increased considerably over time. In central and eastern Europe, where concerted attempts are being made to develop primary care, data from the European Health For All database indicate that to date only Hungary and Latvia have radically reduced the proportion of the health budget allocated to inpatient care. In other countries such as the former Yugoslav Republic of Macedonia and Slovenia, the fraction of health resources allocated to inpatient care has continued to increase (WHO, 2004). Data from western European countries suggests a weak trend towards reducing or containing the proportion of inpatient expenditure in total health expenditure (WHO, 2004). This is consistent with significant attempts to decrease the hospital capacity and to contain hospital costs. At the same time, however, changes in skill mix and technological progress have enabled substitution of primary care for secondary care. OECD Health Data do not suggest a shift of resources towards the primary care sector and outpatient care. In some countries (Belgium, Finland, France, Iceland, Italy, Luxembourg, Netherlands, Spain, Switzerland and Turkey), the proportion of total health care spending allocated to outpatient care throughout the 1990s has remained relatively stable or has decreased. Only two countries have significantly increased the proportion of resources channelled to outpatient care: Austria (6%) and Denmark (3%) (OECD, 2004).

Looking at the distribution of the workforce among levels of care is a way to obtain evidence on possible shifts between primary and secondary care. Available data on the supply of physicians do not suggest increase in the proportion working in primary care. Between countries with referral system (i.e. gatekeeping role) there is also a considerable variation. Within this group, in Norway and Italy the number of GPs per 1000 population is about double the number in Netherlands,

Portugal and Slovenia. As regards nurses, no current data are available to know what proportion are involved in primary care.

Generally speaking, the health system in operation affects both the organization and funding of the primary care sector. Europe's healthcare systems could be divided into two broad groups on the basis of the role played by the state in funding care and making it available to the public. On the one hand are the tax funded models ('Beveridge systems'). In countries that operate such systems, most care facilities are controlled by the government. The EU states with such systems are Denmark, Finland, Italy, Greece, Portugal, Spain, the UK and Sweden. On the other hand are the social health insurance models ('Bismarck systems') in which healthcare is treated as a form of social security and funded by contributions or premiums. Such 'social insurance systems' are in most cases more loosely organised, with the government playing a much more limited role in the provision of care. The EU states with such systems are Austria, Belgium, Germany, France, Ireland, Luxemburg and the Netherlands. In addition to these two groups of West European countries there are the Central and East European countries, where until 1989 the Soviet models prevailed ('Semashko systems') and which have since been migrating to social insurance systems at different speeds.

However, no European health care system is an exact replica of any specific model. Rather, each country has its own variation, in which the basic model is adjusted to national particularities (Marrée and Groenewegen, 1997). As a result also a wide variety of organisational structures for primary care exist in Europe (see Appendix, Tab. 4).

There is some association between the type of system and the degree of access that patients have to elements of that system. In most national health systems, general practitioners have a gatekeeper role; many other care disciplines, such as specialist medical disciplines, nursing and paramedical disciplines, can be accessed only via a general practitioner. Greece and Sweden form exceptions in

this regard: although both countries have tax-funded systems, medical specialists and many other care providers are directly accessible. By contrast, social insurance systems – including those developing in the former Eastern Bloc countries – tend to be characterised by direct access to all components of the system. Again, however, there are exceptions: in the Netherlands, Ireland and Slovenia, for example, access to specialist medical care is generally via a general practitioner. Not all countries with a gatekeeping system apply equally strict rules. This is the case of the Danish tax-funded system. Danes have the option of paying additional contributions to obtain direct access to secondary care, but less than 5 per cent of them choose to do so. In countries with a *referral* or *gatekeeping* system, GPs generally provide a more comprehensive range of services, although they work fewer hours than GPs in countries with equivalent access to medical specialists (Boerma, 2003). Many countries are also examining flexible forms of GP gatekeeping. In countries with gatekeeping GPs, this system is sometimes perceived to be excessively rigid (i.e., in care for the chronically ill), whereas some countries without a gatekeeping system are trying to establish one, initially on a voluntary basis.

Access to hospital emergency departments in urgent cases is provided directly across all countries. However, such departments everywhere find they have to contend with patients presenting non-urgent problems for attention. In the countries that operate a gatekeeping system, this tendency to bypass the gatekeeper is a significant issue. Countries with a GP referral system have been indicated in the appendix (Table 4).

Within Europe, general practices differ extremely in size. According to recently published data, practices in Ireland and the Netherlands are on average three to four times as big as practices in Belgium Finland and France. The average practice size is not related to whether general practitioners play a gatekeeper role. Practice form, on the other hand, is clearly related to the existing type of healthcare system. In countries with social insurance systems, the proportion of single-handed practices is generally well above 50 per cent, whereas under

national health systems the figure is much below 50 per cent. The exceptions are Italy, which has a very high percentage of general practitioners practising on their own despite operating a tax-funded system, and the Netherlands, where there is a social insurance system, but the percentage of practices with only one doctor has been gradually reduced from 66.5 per cent in 1980 to 39 per cent by 2003. Although multidisciplinary primary care teams have been advocate for years, such team formula remain unusual in most EU countries. The exceptions being Finland, where most primary care is provided via large health centres, the UK and, to a lesser extent, the Netherlands. There is a clear link between what a general practitioner is responsible for and whether patient registration and gatekeeping system are in operation. In countries that use patient registration and gatekeeping, a general practitioner is more likely to act as the first point of contact with the healthcare system, both for the general patient population and for people with psychosocial problems.

The type of healthcare system is also closely linked to the mechanisms provided to pay general practitioners. Third payers generally set up how to pay doctors. The nature of the relationship between payers and doctors varies considerably across health systems. Three broad types of relationships can be identified: a *reimbursement approach*, with no formal relationship between the two parties; a *contractual approach*, with a contract between the two parties concerning the method and level of payments; an *integrated approach*, where medical practitioners are employed by insurers (Bickerdyke et al. 2002).

Under the reimbursement approach, third payers receive contributions from the community and use these to reimburse patients for services rendered by GPs (see appendix, Fig. 22) or directly the doctor himself. There is no contracts or employment relationship between third-party insurers and doctors; consumers have freedom to choose their GP; and the payment is typically in the form of fee-for-service (under which general practitioners are paid for the services provided).

This approach can be found in the current Australian Medicare system for medical practitioners.

Under the contractual approach, third payers receive contributions from the community and contract with GPs to make payments for services rendered to patients (see appendix, Fig. 23). The contracts set up the terms and conditions for the provision and payment of medical services and give payers a greater control over funding and its distribution compared with the reimbursement approach. GPs still retain a large degree of autonomy; consumers are free to choose their GP in the public model, but they are generally restricted to contracted providers in the private insurance model; and payment arrangements are usually on a capitation (i.e. a system under which general practitioners receive a fixed amount of money for each patient registered with them) or fee-for service basis, although a salary system (under which general practitioners are paid for the hours they work) can also be used. The public contract system is popular for primary care, playing a part in the UK, Germany, Italy, Ireland, France, the Netherlands and Belgium. Voluntary private contract relationships were the forerunners of Health Maintenance Organisations (HMOs) in the US and used to be common in Europe - but had all but disappeared by the early 1990s (OECD 1992).

Under the integrated system, third payers receive contributions from the community and use these to fund the ownership of primary care premises and to employ general practitioners (see appendix Fig. 24). The third- payers play both the role of insurers and providers within an integrated organisation; consumers' choice of GP in the voluntary private insurance model is likely to be limited and/or restricted by their choice of insurer; and payment arrangements for doctors are usually on a salary basis, although often with the provision for bonuses or productivity payments. Both the public and voluntary private forms have found only limited application in the OECD. In the public model, which is dominant in Spain and formerly used in the United Kingdom, the government is both the major insurer and provider. The voluntary private integrated system, on the other hand,

is characterised by private insurers and forms the basis of the current US 'managed care' system (based originally on (HMOs), but currently with a variety of forms). Most of the working age population in the US is currently enrolled in either HMOs or their equivalents (Rice 2002). Under this model, HMOs receive a fixed amount from the population (often employers) for providing services to enrolled patients - with the payment unrelated to how much the HMO actually spends. While patients are covered by their insurance they will usually pay a co-payment each time they visit their doctor.

Most national health systems pay general practitioners on the basis of a capitation system, The capitation system operated in the UK until the New General Medical Services Contract was introduced, and still operates in Italy as well as in the Netherlands (where the obligatory insured majority of the population is concerned) and in Belgium's (community) health centres. Finland, Greece, Portugal and Sweden all have salary systems. In Denmark, half a general practitioner's income is accounted for by age-related capitation fees and the other half by payments for particular services. In several countries of central and eastern Europe, the salary system dominates. However, in Slovakia, the great majority of the general practitioners are self-employed and are paid on a fee-for-service basis. In the Czech Republic, a combined capitation and fee for service system has been established since1998, in which the capitation payment accounts for 80per cent of the practice income. To a lesser extent, this is also true of the Czech Republic.

The prevailing payment system also influences how GPs respond to varying workloads. In countries with *self-employed* GPs paid on a fee-for-service basis, the GPs are more concerned with the treatment and follow-up of diseases and they devote more of their working hours to direct patient care than GPs in other countries. *Salaried* GPs provide fewer treatment services than self-employed GPs. Services like preventive screening, which are not 'demand-driven', are unlikely to be provided under simple capitation payment systems, which means that additional target payments are required.

In all EU countries where capitation payments account for an important part of the funding given to general practitioners – i.e. Denmark, Ireland (where lower income groups are concerned), Italy, the Netherlands and the UK – everyone has to register with a particular general practitioner, group of general practitioners or health centre. Portugal, Spain and Slovenia also run patient list systems. In principle the patient list system is not directly linked to the gatekeeping role. Gatekeeping can be found in countries where patients are not registered with a primary care practice, such as Iceland, Croatia and, until introduction of the patient list system in June 2001, Norway. Within the European Union, however, patient list systems are always found in conjunction with gatekeeping arrangements: all EU countries either have both or have neither. Though there are no comparative international data on this issue, it seems quite plausible that registration with primary care facilities strongly contributes to the realisation of continuity and coordination in the delivery of healthcare. With respect to this last aspect, a system in which patients are registered with a GP offers a greater likelihood that medical information is stored in one place, than do systems without patient lists. A patient list system is not sufficient, however. Individual GPs need to keep comprehensive medical records and keep good working relationships with other health professionals in primary and secondary care. Computerized medical records are not just helpful for coordination and in providing continuous care to individual patients. A good practice database is also essential for the systematic screening and follow-up of patients affected by chronic ills. Routinely kept medical records become a major source of information for both epidemiological and health services research (Rosser and Van Weel, 2004; see also chapter I and III in this thesis). In some countries, the patient list system also applies to pharmacies. Registration of patients in one pharmacy greatly facilitates an active role of pharmacists in primary care, such as monitoring the medication of individual patients.

The UK has a system for funding primary care that is unique in the EU. In the National Health Service, remuneration used to be based mainly on capitation, supplemented by fees for certain specific services and for achieving certain target

levels of service (involving only childhood vaccination and cervical cytology). In the 1990s a fundholding scheme was introduced. General practitioner fundholding was a form of integrated capitation, i.e. a system associated mainly with Health Maintenance Organizations in the USA, under which the services provided by various care providers or at various levels of the healthcare system are paid for out of a single general budget. The UK's fundholding scheme allowed general practitioners to buy hospital care for their patients. With effect from April 2004, however, the UK has introduced the New General Medical Services Contract. Under this contract, each general practice receives a basic sum based on the size and make-up of its patient list, to cover the cost of providing basic family doctor care. Additional forms of care that general practitioners are not obliged to provide, such as more complex minor surgery, mother-and-baby consultation services, support for drug-users and the homeless and out-of-hours care, are paid for separately. A practice can also qualify for additional payments if it realises certain quality standards, measured by reference to a total of 136 indicators relating to medical treatment, practice organisation and patient-orientation.

The social insurance systems of Belgium, Germany, France and Luxemburg have fee-for-service arrangements, under which general practitioners are paid a certain amount for each type of service they provide. Other countries, such as Ireland and the Netherlands operate mixed payment systems, with general practitioners paid a fixed capitation fee for some of their patients and service-related fees for the rest. In Austria, some *Bundesländer* have a fee-for-service system, while others pay capitation fees.

In primary care, medication is normally funded separately, in contrast to the situation in hospitals, where medicines are usually paid for out of the institution's overall budget. Here again, there is considerable variation within the EU, which is not related in any way to the type of healthcare system in operation. The arrangements vary from full out-of-pocket payment (in Lithuania), fixed fees (in the UK, Germany and Austria), coinsurance with patient charges (in Belgium, Greece, Luxemburg, Poland, Portugal and Slovenia), co-insurance without patient

charges (in reimbursement (in the Netherlands, Slovakia and Italy) and full reimbursement of only the cheapest generics and considerable co-payment of other medication (in the Czech Republic).

In the more densely populated parts of the EU, the geographical accessibility of general practitioners is not normally a problem. In more remote rural areas, however, accessibility can be compromised by a shortage of general practitioners. In the Scandinavian countries, the Netherlands and the UK, one normally needs an appointment to see a general practitioner. In Finland and Sweden, people often have to wait two or more days for a consultation. In Hungary, Italy and Latvia, by contrast, general practices with appointment systems are unusual. It is usually possible to contact one's local practice by phone, but the frequency of telephone consultations varies from country to country, from two to sixteen per general practitioner per day. There is even greater variation in the number of house calls made by general practitioners. In Portugal, Sweden and Finland the average general practitioner makes two or three house calls a week, while in Germany the number is thirty-four a week, and in Belgium forty-four. Generally speaking, accessibility is even more closely related to population density in other primary disciplines than it is in general practice. In remote rural areas, the general practitioner is often the only available care provider, apart from the pharmacist, who is generally accessible for anyone during office hours. In some countries, including Norway, Switzerland and the Netherlands, general practitioners in remote areas sometimes run their own pharmacies.

Throughout the EU, primary care is generally available outside office hours in emergencies. In many countries, the general practitioners and pharmacists in a given area operate a cooperative out-of-hours scheme, covering for one another on a rota basis. In Italy, however, care is provided outside normal surgery hours by a separate government run locum service. In 1992, Denmark switched from locally organised rota systems for out-of-hours primary care to county-wide locum schemes each covering fifty to sixty thousand people. This resulted in a considerable reduction in general workloads. In the UK, 'NHS Direct' has been in

operation since 1998. NHS Direct is a twenty-four-hour telephone support service staffed by trained nurses who provide advice and triage. In the Netherlands, a network of general practitioner stations has quickly sprung up since the late 1990s at the instigation of the general practitioners. At these stations, practice assistants or nurses working under the supervision of general practitioners field requests for general assistance outside surgery hours. Each station is responsible for a region with an average population of 150,000.

The density of doctors seems to be linked to the growth of the nursing role: in Italy and Germany there is a high density of doctors and a rather limited nursing role, while in the UK and Canada there is a lower proportion of doctors and the nursing role is more developed. However, a modest density of doctors is not automatically related to the development of the nursing role in primary care. In countries such as Germany and the Netherlands, other professionals (medical assistants) working with doctors in particular, enable group practices to extend the range of services offered. It seems that close collaboration between GPs and nurses is not a natural form of working in most countries. In fact, where this model does exist, whether in an experimental form (Italy, Ontario, Quebec, the Netherlands, Germany), or more generally (the UK, Sweden, Finland), this is as a result of concerted policy implemented by the government. It has been demonstrated that practices with more staff and equipment provide a wider range of services. Moreover, GPs working in groups may be more efficient because they work fewer hours with similar workloads. The density of GPs across Europe and some extra-UE countries is shown in Tab. 5 in the appendix. The table shows considerable variations across countries. Comparison of the figures from 1990 and 2002 shows that the numbers have remained relatively stable in most countries. The Italian health system is unique in that it has a surplus of doctors.

Recruitment for general practice is influenced by the expected income earned as a GP compared to other medical specialists. Physician incomes throughout most developed countries are very high, among the highest for any occupational group. Differences in income reflect differences in status between medical specialties.

Existing data suggest that physicians' incomes, both generalists and specialists, are in general in the top 25 per cent of the population, similar to senior civil servants (OECD, 2000; Reinhardt *et al.*, 2002). However, there is huge variation among countries. Doctors in western Europe receive higher remuneration than their colleagues in central and eastern Europe. GPs' incomes have been increasing in many countries over recent years, but data available for a few OECD countries show a gap between average physicians' income (specialists and generalists) and average GPs' income in most of these countries (OECD, 2000). In some countries such as Norway, systematic attempts have been made to reduce the income gap between generalists and medical specialists (Furuholmen and Magnussen, 2000). In all countries, the average income among physicians was superior to the average income among other medical professionals (OECD, 2000). In general, physicians' average wages rise more than twice as high as wages for nursing and midwifery professionals (Gupta *et al.*, 2002). There are no specific data on primary care nurses.

In Italy primary care is provided by general practitioners or paediatricians working under a government contract who are paid a capitation fee based on the number of people (adults or children) on their list. Occasional visits to patients not included in the list are paid directly by the patient (the fee is defined for ambulatory and domiciliary visit). The same fees apply to foreign patients who temporally reside in Italy. Primary care is provided by 47100 independent contracted doctors (one for every 1059 inhabitants), the equivalent of the British GP. Primary medical care for children is provided by 7100 pediatricians (one for every 1075 children under 14 years).

GPs are responsible for referring patients to secondary and tertiary care and could, in principle, serve a gatekeeper function. They write prescribing diagnostic interventions, drugs and certifications and also visit patients at home if necessary. Primary care physicians are authorized to work in the National Health System (NHS) after successfully completing a 2-year specialization course in general medicine and acquiring clinical experience as temporary staff in NHS facilities.

The vast majority of GPs work in single practices, although both national contracts and regional agreements try to encourage group practice by offering supplements to the capitation payment, extra resources for the practice, including secretarial and nursing support and basic diagnostic equipment

Patients are free to choose their doctor provided that their list size does not exceed the maximum permitted (1800 for a general practitioner and 1000 for a paediatrician). In 2003, each general practitioner averaged 1099 patients (HFA, 2005) The regional range of averages ranged from 994 (Lazio) to 1344 (Trentino A. Adige). The percentage of general practitioners with a list size greater then 15,000 patients ranged between 5.12 (Liguria) and 37.26 (Trentino A. Adige). The national average density of general practitioners was 8.18. See Tab. **6** for more details.

GP remuneration and work conditions are regulated by a national contract, complemented by regional agreements. The National Contract is a result of negotiations between the government and representatives of general practitioners organized in various trade unions. Once reached, the content of the Nation al Contract is legislated through a Decree approved by the Ministry of Health (i.e., the agreement is a binding by law).

According to the National Contract for General Practitioners, primary care services are reimbursed through three main channels: 1) per capita fee (age-adjusted), negotiated and fixed at national level; 2) extra funding linked to performance targets set at national/regional/organizational level; 3) fee for service for additional services defined by the contract. Per capita fee is a fixed part of GPs funding. In addition to fixed amount per patient, GP receives funding in relation to their age, years of experience, and number of patients enrolled in the list (per capita fee significantly decreases with higher number of patients). There is no empirical evidence of any kind of cost assessment conducted in order to define these fees.

Recent reforms introduced extra funding to encourage group practices - which are considered the base from which the range of primary care services may be enlarged - and to promote integration between primary care physicians and services directly offered by health authorities such as specialized medicine, social care, home care, health education and environmental health. In addition, extra funding is available for upgrading the quality of service provision by improving, for example, physicians' practices information systems. It is estimated that up to 10% of GP income comes from extra payments for participation in special programmes or reaching organizational or expenditure targets.

Specifically, in Lombardy some experiments with 'primary care groups' have been established with a view to improving the follow up of diabetic and hypertensive patients. Follow-up protocols are implemented by nurses in the doctors' practice, under their responsibility. These nurses do not make diagnoses or prescribe (Yann Bourgueil, Anna Marek, Julien Mousquès, 2005). The national contract which regulates relations between doctors and health authorities states that the salary of a nurse should be 35% of that of a doctor practising full-time with a list of 1500 patients. This approach was also chosen by Emilia-Romagna which focus on strengthening primary care. Its regional policies aim to promote group practices, and to integrate general practitioners, social services and public health, and to redistribute among health professionals. Incentives for reorganisation include in particular, help with finding jobs for personnel (secretaries, nurses etc.).

Finally, a special section of the National Contract (Appendix D) defines "additional services" that can be provided by General Practitioners and for which specific tariffs are applied. For their provision GPs are paid individually, on top of their basic remuneration. These services are classified in three categories: (i) services not requiring NHS authorization (first and subsequent medication, superficial wounds suture, threads removal, urethral catheterization (men and women), phleboclysis (only in emergency cases), tetanus vaccination); (ii) services requiring NHS authorization (phleboclysis cycle curative cycle of endo-venous injections, aerosol curative cycle, non obligatory vaccination) and (iii)

services defined by additional regional agreements for residents, non residents and foreigners.

The categories of services that primary care physicians are obliged to provide under the National Contract signed on January 28th 2005 to ensure 'to guarantee the provision of of essential levels of care across the national territory' are defined broadly as: (a) essential services: acute and chronic disease management, in line with best practice indications and in agreement with the patient; (b) health promotion activities; (c) patient management with in programmed and integrative domiciliary care coordinated with providers of specialist and rehabilitative care services; and (d) community services defined on the basis of region al agreements.

Regions are autonomous in establishing further agreements (*Accordi Integrativi Regionali*) aimed mainly at identifying the most appropriate organizational arrangements for the provision of services set at national level. The National Contract also encourages various forms of integration between primary care physicians and district services such as social and home care. Additionally, the National Contract obliges Local Health Units to guarantee continuity of care, i.e., primary care services 24 h a day, 7 days a week. Organizational arrangements are decided at regional level. Finally, they set specific conditions for extra funding.

Thus, it is clear that the organization of a health care system, and more generally, the institutional setting are an important determinant of the way in which primary care services are provided (for more details on the reforms of health care system see the extra section on the Italian Legislation, in the appendix – Box 1.1). However, due to a lack of data, the picture that may be sketched of primary care in the EU, is far from complete.

CHAPTER I

QUALITY INCENTIVES IN PRIMARY CARE - THEORY

1.1 Introduction

The incentive to provide quality in primary care has received increased recognition by policy-makers and researchers, especially over the last decades. Specifically, policy-makers are interested in guaranteeing the provision of a high quality service - to a patient or a population - and in containing the health expenditure. Bringing together both of these objectives inevitably lead to a trade-off. As for research, the issues concerned with quality of primary health care and its incentives come not only from the health economics literature but also from other disciplines, including: industrial organisation; regulatory economics; theory of incentives and mechanism design; and managerial economics; as well as combination of sociology and economics.

The aim of the chapter is to draw on the contributions of several researchers to develop a working framework to model the quality-related physician behaviour. It is built in such a way that it develops this framework stepwise, starting with models of complete information where non-retradability of medical services and non-contractible input (quality or effort) lead profit-maximizing physicians to influence the quantity demanded by patients. Subsequently, the third way of influencing patient demand is discussed: persuasion. Persuasion can be used by physicians in order to increase demand and thus maximise income. In that light,

explicit attention is paid to physician-induced demand (PID), one of the most discussed and examined issues in health economics. Attention is then paid to the alternative objective of a target income. The target income hypothesis is often used to explain PID (i.e. demand may be induced in order to ensure an acceptable income-level for the physician)[8].

The discussion then focuses on the different forms of incentives, namely competition, regulation, and values and norms. Incentives are distinguished between monetary and non-monetary, the former arising from payment system and the latter from regulatory control as well as from ethical and professional norms.

Finally, we move away from the pure profit maximization paradigm and discuss other physician objectives, such as power, motivation, ethics, and altruism.

1.2 Defining and measuring quality in primary care

There is no universally accepted definition of quality of care. It has been defined in a number of different ways (Crosby, 1979; Donabedian, 1980; Maxwell, 1984; Peters, 1987; Juran, 1988). Definitions of quality are either generic or disaggregated. These approaches are not inherently incompatible but can be seen as opposite ends of a continuum. Generic definitions of quality include excellence (RCGP, 1994), expectations or goals which have been met (Ellis and Whittingham, 1993; Steffen, 1988), `zero defects' (Crosby, 1979) or fitness for use (Juran, 1988). Other generic definitions are more complex. For example, the Institute of Medicine (Lohr, 1992) has defined quality as the `*degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge*'. Generic definitions are not easily operationalisable and trade both sensitivity and specificity for generalisability. Disaggregated approaches, on the other hand, recognise that quality is complex and multidimensional (Donabedian, 1980;

[8] McGuire (2000) argues that the target income hypothesis stems from the behavioural economic stream of literature, but subsequently tries to formalise it in a utility maximising framework.

Maxwell, 1984; HSRG, 1992; Winefield, Murrell & Clifford, 1995). They define quality according to individual dimensions or components. Quality definitions also vary by country, disease burdens, resource constraints, and by a particular health system's level of development. In the USA, for example, adequately stocked drug inventories are presumed in all hospitals and would not typically be mentioned as an indicator of quality, whereas in many developing countries, this is not the case. For example, in some cultures, good quality means that an adequate number of caregivers staff the office or facility. In others, good quality means optimum clinical outcomes. Different stakeholders attach different levels of importance to different aspects of quality. From the point of view of consumers for example quick attendance to patients and availability of drugs are noted as important attributes. They can also take into account the physician technical competence, such as knowledge levels for assessing and managing common and serious conditions. Thousands of measures have been developed to quantify and compare health care quality, but there is no consensus on which subset is most appropriate to use as a barometer of quality. Quality measurement itself can be daunting, particularly in health systems with emerging information and monitoring systems.

Campbell et al., (2000) propose a combined generic and disaggregated approach to defining quality of care. Their definition of quality of care for individuals is 'whether individuals can access the health structures and processes of care which they need and whether the care received is effective' (p. 1614). From this definition it is possible to identify two main components of quality at a micro level, namely access and effectiveness. Access must be intended in terms of both geographic/physical access (Haynes, 1991) – the extent to which there exist barriers such as rurality, disability or old age to reaching health care facilities - and of availability - the extent to which the health care system provides facilities (structures) and services (process) to meet the needs of individuals. Effectiveness, instead, must refer to both clinical and interpersonal care. The first deals with effective application of evidence-based or legitimate medicine while the second takes into account the fact that processes and outcomes depend on the specific

characteristics of the patient. Coordination or integration of care for individual patients is also recognized as another aspect of effectiveness and is a further attribute of particular relevance to primary care (Starfield, 1994). By contrast, on the macro-side what matters is 'the ability to access effective care on an efficient and equitable basis for the optimization of health benefit/well being for the whole population' (Campbell et al., 2000, p. 1617). Accordingly, equity and efficiency are the additional components of quality for population. Conflicts between the micro-efficiency an the macro-efficiency aims can arise, leading to trade-offs between high quality and cost-containment on one hand, and efficiency and equity on the other (Kuhn, 2003). Starting from this framework, several studies tried to develop and apply indicators with the aim to reflect different dimensions of quality in general practice from diverse perspectives, using different methods and sources of data. However, a discrepancy between the definition of the multidimensional aspects of care and the methods currently used for collecting data to measure such aspects persists. For example, data contained in administrative datasets used for financial purposes cannot be applied to obtain a comprehensive quality assessment but for assessing those aspects of quality 'which are consistently and reliably recorded therein' (Campbell et al., 2000). Campbell et al. (2000) suggested using this framework to permit critical analysis of quality indicators and to show which aspects of quality they measure and which ones do not.

1.2.1 Operational definition of primary care quality

Once a definition of quality is determined, it must fit into an operational framework so that it can be measured, changes may be implemented, and then evaluated. Measuring the quality of care has traditionally relied on a framework developed by Donabedian (1980) of structure-process-outcome, which is outlined below. This thesis embraces and applies this definition for quality in primary care.

Structure

Structure refers to the resources available in the primary health sector. Resources may be categorized as individual and group practitioners (their attributes such as

age, qualification, skills, specialty board certification, licensure, type and level of training), as well as facilities (location, ownership, patient load, accessibility, government certification and accreditation, physical attributes, including safety, policies and procedures) (Friedman, 1995 and Donaldson, 1999) and medical capital. Structure asks the question, "are the necessary resources available to provide effective, efficient medical care?" (Longo, 1994). The problem with evaluating with structure variables is that there is substantial evidence that the relationship between structure variables and process and outcome variables are weak, inconsistent and sometimes paradoxical. Developing public information on quality of care based uniquely on structural measures can be misleading (Meyer, 2001).

Process

Process measures the performance of physicians, and can include aspects of how consumers seek and obtain care and their interaction with physicians. Physician performance can be based on interpersonal care (service, timeliness and convenience) and on technical aspects (timeliness and accuracy of diagnosis, clinical examinations, appropriateness of therapy and treatment, and prescriptions) (Friedman, 1995 and Donaldson, 1999). Process quality derives from the variable input choices taken by the physician and his own effort. Greater effort improves the quality of process. Compared with skills, effort can be chosen freely and adjusted in the short-term (Kuhn, 2003). Process asks the questions, "Have the processes necessary for providing effective and efficient medical care been provided?" and "Are these processes in control, ensuring that 'outputs' consistently meet requirements?" (Longo, 1994). For example, a process measure would ask if patient with diabetes have had their eyes and kidneys tested at least once every two years.

Process quality is the harshest judgment of the quality of care. For the vast majority of medical conditions process measures will need to be used to assess quality. Regardless of what we would like to have happen, most of the quality indicators that we should use will be process based (Brook et al., 2000). Process measure are most valid when they have been shown to directly improve health.

Outcome

Outcome is the end result of care, or what has happened to patients, including measures of survival, unintended effects of treatment, and the relief of symptoms (Donaldson, 1999). End results include: health status, functional status, mental status and the general well being of the patients and populations (Friedman, 1995; Brook, 2000) Outcomes can be measured in quantifiable terms, such as morbidity and mortality, and with qualitative measures, such as patient satisfaction. One of most significant problems with health outcomes is that they are affected by uncertainty. There might be other factors involved with changes in health outcomes besides quality improvements, resulting in weak association between health care and outcomes. In addition, changes in health outcomes of interest often occur several years later. There are also problems of differences in case mix. The same disease may have different levels of severity, risk adjustment is difficult and costly, and sometime is impossible due to unavailability of data (Brook et al., 2000). Positive outcomes for an asthma patient may include the ability to function normally, play sports or avoid hospitalizations (www.healthjournalism.org/qualityguide/chapter1.html).

Quality may be measured in each of the three components or in some combination. The literature contains conflicting views about which measures are most useful and productive, and what should be measured at what time. There has been considerable debate about whether process or outcome should be assessed as measures of quality of care (Davies & Crombie, 1995; Brook, McGlynn & Cleary, 1996). In general, process measures are better indicators of quality of care if the purpose of measurement is to influence the behaviour of the health care system: processes are common, under the control of health professionals, and may more rapidly be altered. Outcomes are often rare, may follow a change in process by up to 10 years (e.g. management of hypertension), and may be dependent on factors outside the control of the individual health professional (Giuffrida, Gravelle & Roland, 1999). However, process measures suitable for measuring quality should be clearly linked to evidence of improved outcomes. To this purpose we tried to make this link clear (see chapter III).

Physician can influence the quality of structure by making appropriate investments and the quality of process by exerting effort (see Fig. 1).

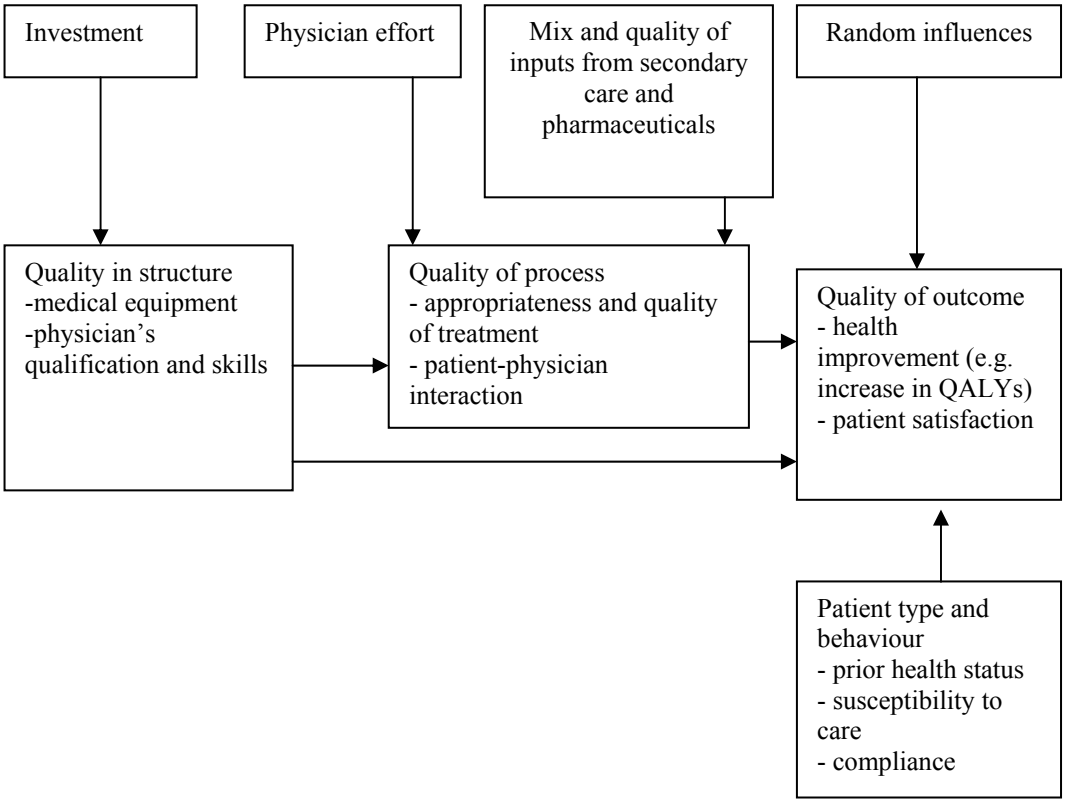


Fig. 1 Schematic representation of quality in the production of primary care (Source: Kuhn, 2003).

The choice of the level of investments and effort are crucially determined by the incentive system and the resource constraints (see Fig. 2). Incentives can be distinguished between monetary and non-monetary, the former arising from payment system and the latter relating to ethical and professional norms as well as from regulatory control. In maximising their objectives physicians are subjects to resource constraints, which place a bound on the provision of quality. The level of quality achievable given these resource constraints depends both on incentives and on the institutional context. Institutions play also a role in determining the

quality of care. The term 'institution' refers to both the organisational context (e.g. whether the physicians work as individual or group practitioners) as well as to the formal and informal rules which are applied (e.g. the payment system, the presence of standards of care or of social norms). Policy-makers shape the quality incentives by designing institutions. In doing this, they face a trade-off quality against cost-containment and efficiency against equity.

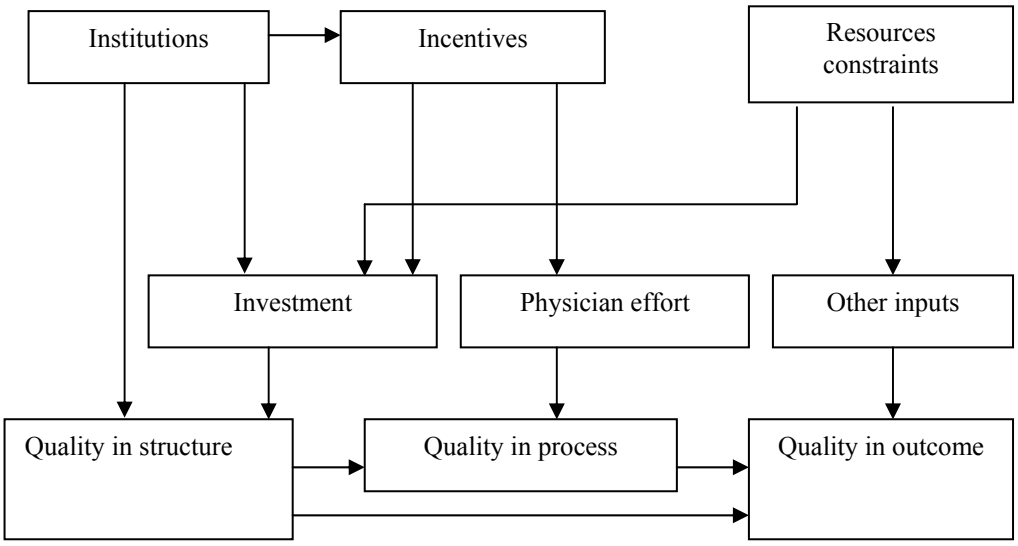


Fig. 2 Roles of institutions, incentives and resource constraints (Source: Kuhn 2003)

In the reminder of this chapter we focus on the definition of quality indicators as developed in the literature and, then, we move on to discuss the physician's decision making process focusing on the role of incentives in stimulating quality.

1.3 Rationale for measuring quality in primary care

Developing and applying quality indicators have become a significant field of research in primary care assessment over the last thirty years. The rationale for measuring quality through indicators relies on the necessity of detecting problems (such as unjustified variations in primary care) in order to outline an adequate

quality improvement strategy[9]. A range of indicators can be developed depending on the stakeholder's perspective (patients, cares, managers, professional, third party payers); on the aspects of care to assess (structure, processes or outcomes); and on the methods used (non-systematic or systematic) (Campbell et al., 2002). Several studies documented variations in quality of primary care applying indicators derived from either secondary data source (namely, routinely collected hospital episodes) or primary data source (namely, clinical audit). Among them, one strand suggested the central role for primary care in preventing unnecessary hospitalizations while another more recent strand focused on measuring areas other than the clinical one to include different aspects of quality such as access, interpersonal care, continuity of care (Campbell et al., 2001). It seems that only a few studies have investigated the relationship between measures of quality derived from these two approaches (Reid et al, 1999) and no studies comparing preventable hospitalizations with actual practice quality (clinical audit) were found. The strengths and limitations of using such indicators developed from different sources of data are illustrated in the following sections and some evidence on their application is given in chapter II.

1.3.1 Outcome versus process quality indicators

This section focus on two categories of indicators used in the literature for the assessment of quality in the primary care delivery system. One set of indicators is derived from variables provided by routinely selected hospital episode data, namely outcome indicators; the other is obtained from GP practice records according to explicit criteria, namely process indicators. The first category of indicators is based on a secondary data source while the second one is based on a primary data source. While indicators derived from routine data take into account the clinical outcome related to the preventive care and the management of chronic

[9] Investigating variation involves both inefficiency and inequity issues. Inefficiency arises when the health gain is not maximised, i.e. the marginal patients in high use areas have lower benefit than marginal patients in low use areas. Inequity arises when similar patients are treated differently. To distinguish variation from inequity it is important to know how much of variation is: justified (due to morbidity, distance, waiting times, age, gender, etc.), unjustified (age, gender, ethnicity budgetary arrangements) and unexplained (Ferguson et al. 2002).

disease, the latter also allow for measuring aspects of quality other than the clinical one, namely access, effectiveness and interpersonal care. This last set of indicators has represented the ground for the most comprehensive assessment of quality of care in general practice in England (Campbell et al., 2001).

1.3.2 Quality indicators based on routinely collected hospital episode data

In several studies on quality researchers have used secondary data because these are readily and quickly available and at the same time provide data for large time periods allowing for a retrospective use. In addition, the cost for their collection is low (being often assembled for other purposes such as hospital reimbursement), especially if compared to the large amount of information provided about several patients with different diseases across different geographical and healthcare scenarios. They also involve less ethical and consent permissions than the gathering of data through a primary source. It is clear that the appeal of this approach derives from the fact that the majority of countries have good data on hospital admissions/discharges and limited data about ambulatory care. However, their accuracy and completeness may vary (Powell et al, 2003). It is recommended to use valid and reliable outcome measures to assess quality when using these data, being aware that applying different measures to assess the same practice could lead to conflicting findings. Compared with process indicators, most of measure of outcomes are relatively immune to manipulation by providers (although physicians may be able to influence the risk-adjusted outcome measures by exaggerating the risk characteristics of their patients). They also are long-term in nature, encouraging providers to adopt technologies that recognize long-term benefits (Goddard et al., 2002a; 2002b). Finally, focusing on outcome measures lets to achieve a single goal, namely the patient's health status, rather then specific interventions.

Hospital admissions for ambulatory care-sensitive conditions (ACSCs) is a commonly used outcome indicator for measuring access and effectiveness in

primary care derived from such data. Ambulatory care-sensitive conditions were derived from the sentinel health conditions as proposed by Rutstein et al., 1970) [10] but they are more broad than the sentinel condition model. They refer to diseases that are sensitive to outpatient care. Examples are given by vaccine-preventable diseases, early recognition and excision of melanoma, and effective glycaemic control in diabetics (Jackson & Tobias, 2001). For these events, a timely and effective outpatient intervention is likely to reduce the risk of hospitalization by 'preventing the onset of the illness or condition, controlling the acute episodic illness or condition, or managing the chronic disease or condition', respectively (Billing & Hasselblad, 1989). ACSCs are also called potentially avoidable hospitalizations (PAH)[11]. They need to be distinguished from population preventable hospitalizations resulting from diseases preventable through population-based strategy (e.g. tobacco excises or smoke-free laws) and from those hospitalizations avoidable through injury prevention (wearing of seats belts, domestic hot water temperature reduction, smoke alarms, etc.)[12].

Giuffrida et al. (1999) pointed out that crude rates of adverse events – in this case the ACSC admission rates - give only an indication of the size of the medical problem in a population, that is the health outcome. A valid performance indicator should reflect only those aspects of care that are under the control of the staff (and the health care system and organisation in general) and able to improve that

[10]Broadly speaking, sentinel conditions are alarm bells for problems with the personal or public health system. Three sets of sentinel events are identified: 1. events that occur because of complications due to late treatment (Schreiber & Zielinski, 1997) 2. singular conditions which should never occur, e.g. child with polio and 3. conditions which cannot be totally prevented but the incidence of which can be reduced, e.g. low birth weight babies (Arnold & Zuvekas, 1998).

[11] The most update list of diagnoses (ICD-9-CM ICD-10-AM class) that identify ACSC include Flu, Chronic Obstructive Pulmonary Disease and Ear-Nose-Throat; Vaccine; Asthma; Congestive Heart Failure and Hypertension; Diabetes; Angina; Anaemia; Nutrition; Dehydration; Gastroenteritis; Pyelonephritis; Ulcer; Cellulitis; Pelvic Inflammation; Epilepsy; Gangrene; Acute Appendicitis; Hernia; Thyroid; Rheumatic Fever; Skin Cancer; Breast Cancer; Cervical Cancer; and Stroke (Jackson and Tobias, 2001; The Victorian Ambulatory Care Sensitive Conditions Study, 2001–02, 2004). However, some inconsistencies there exist in classifying ACSCs across a number of studies.

[12] McPhillips et al. (2001) demonstrated that infants hospitalized before 90 days of age are an easily identified and potentially target able group of infants for injury prevention interventions. Home visits by nurses has been shown to be effective in reducing injury hospitalizations in other high risk groups of infants, and may be an effective intervention in infants discharged from the hospital in early infancy, particularly if lack of community resources prohibits intensive home visit programs for all infants with high risk environments.

outcome/access. They documented that between a third and a half of the variation in admission rates across health authority areas depends on confounding factors. For this reason, it is essential that the measures obtained can be retrospectively adjusted for in the analysis[13]. Therefore, 'indicators' such as admission rates (in fact, health outcomes) cannot be used directly as measures of the quality of primary care. To which extent do crude admission rates relate with quality needs to be assessed since the relationship is not immediate.

Another potential problem of using admission rates as indicators of quality of care refer to the role of the statistical instability of any relatively rare event (Goddard, 2002b). As a consequence, the indicators may not be applicable to individual practices, where greater year to year variation would be expected to occur because of the smaller population size. Instead, this problem can be reduced for a health authority by using a moving average model (Giuffrida et al., 1999; Powell et al., 2003). Moreover, when developing rates of admission for ambulatory care sensitive conditions, two further issues needs to be taken into account. One is associated with the multiple admissions (i.e. the hospitalization of the same person for more than one event during a year which may affect the independence assumption on events) and the other with border crossing (Ricketts, 2001). In particular, multiple admissions results in an overestimate of hospitalization for a given condition while when primary care access area crosses the borders this may affect the computation of the admission rate for a given area. Such data are also limited by the ecological problem: the assignment of community-level characteristics as a proxy for individual ones can decrease or, more frequently, increase estimates of individual-level associations (Morgenstern, 1982; Lambrew, Carey and Billings, 1992). Casanova and Starfield (1995) reported that some authors disagree (Gonnella, 1977; Krieger, 1992; Morgan et al, 1983; Curtis, 1990). It is also worth to notice that an implicit assumption when using ACSC to make comparison between different practices to investigate variations in the quality they delivery is that the prevalence of the underlying condition should be

[13] This includes a risk adjustment and controlling for the other relevant patient-specific characteristics, including both clinical and non-clinical factors (such as age, co-morbidities, past medical history, socio-demographic factors), as well as for contextual factors (such as the supply of secondary care resources or the policies of local specialists). All these variables might influence hospital admissions but be beyond the control of the GP practice.

similar between the units of the analysis, or should be estimated, so that it can be adjusted for in analysis[14]. In addition to prevalence even co-morbidity factors need to be taken into account being important predictors of outcomes. These include factors such as mental disorders, drug and alcohol abuse, obesity, etc. Co-morbidities may be associated with substantial increases in length of stay, hospital charges, and mortality. Finally, the meaning of ACS may vary between urban and rural areas (Schreiber and Zielinsky, 1997).

1.3.3 Indicators derived from GP records.

Directly measuring the adequacy of service in terms of type, timeliness, intensity and location (Restuccia et al., 1989) through a survey of the population of interest is the alternative to using (adjusted) ACSC hospitalizations rates as indicators and also the most logical approach. However, the direct measurement of adequacy is resource-intensive and generally impractical (Schreiber and Zielinski, 1997). In England a survey of 60 general practices was carried out in 1998 providing detailed measures of quality related not only to the management of chronic diseases and preventive care but also to prescribing, access to care, continuity of care, and interpersonal care derived from clinical audit[15]. However, data abstraction from records have been found to underestimate quality of care because records are not able to catch what characterises preventive or counselling/advise activities (Luck et al, 2000). There exist some gap between what physicians do (observed procedures) and what they record. Therefore, poor audit results can either reflect poor care or poor recording. Some authors (Solomon et al. (2000); Kosecoff et al. (1987)) found that quality of record keeping is positively correlated with quality of care in the US system. However, in UK this kind of studies has not been carried out yet (Campbell et al., 2001). However, when poor performance on a process measure reflects poor quality, this gives a clear

[14] This paper will tries to address this issue among the others.

[15] A three stage process to select practices was used. Three out of the eight English NHS regions were selected - North Thames, North West, and South West - as being nationally representative in terms of rurality, socioeconomic deprivation, and geographical dispersion of population. From each of these three regions two health authorities was selected as being representative of their region in terms of rurality and socioeconomic deprivation. The six health authorities selected were Bury and Rochdale, West Pennine, Enfield and Haringay, South Essex, Avon, and Somerset. Finally, within each of these six authorities a random sample of 10 practices stratified in terms of practice size, training status, and socioeconomic deprivation was selected. These 60 practices were invited to take part in a detailed assessment of quality. When a practice refused to participate, another with similar characteristics was chosen at random and invited to participate; 60 out of 75 (80%) practices agreed to take part. **(**Campbell, 2001 BMJ).

suggestion of the remedy required. By contract, poor performance on an outcome measure gives no such indication (Goddard et al., 2002a).

Review criteria applied to assess quality in delivering primary care are selected from clinical audit by using systematic methods that combine high quality research evidence with expert opinions. Using review criteria should ensure that not only changes in the structure but also (and above all) improvements in outcomes of primary care can follow from improvements in quality of process (Hearnshaw, 2003). By contrast, assessing quality against inappropriate review criteria can lead to wasting of resources for ineffective quality improvement activities. The list of items in the Appendix to Chapter III (Tab. 8; Tab. 9; Tab. 10) provides a group of review criteria that were used by primary care organizations and general practitioners for assessment of the quality of care they deliver to patients with chronic diseases (angina, adult asthma, and type 2 diabetes)[16]. They were developed in 1997. Within primary care there has been also a particular focus on indicators for prescribing[17], on prevention care indicators[18], and on care delivered[19].

The literature review on quality indicators has tried to evaluate outcome and process indicators against their ability to measure quality in the primary care.

[16] To assess the quality of clinical care computerised disease registers or prescribing records was used to select 20 patients in each practice receiving maintenance treatment for each of three conditions: asthma in adults, angina, and type 2 diabetes mellitus. After confirming the relevant diagnosis from the medical records, data from medical records was extracted to identify aspects of care previously defined by expert panels as being both necessary to undertake and necessary to record for these conditions (Campbell et al., 1999). These criteria are listed in Appendix 3.

[17] Campbell et al. (2000) by using a Delphi consultation study identified four indicators for UK general practice rated valid and reliable to assess quality: 1. ratio of bendrofluazide 2.5 mg items to all bendrofluazide items; 2. % of antibiotic items contained in predefined list (health authority, primary care group, or practice formulary); 3. items/STAR-PU for antibiotics, DDDs benzodiazepines/benzodiazepine STAR-PU (including zopiclone and zolpidem); 4. ratio of co-trimoxazole items to trimethoprim items. However, these indicators have only a narrow focus allowing a limited interpretation of the quality of prescribing among a group of GPs.

[18] Preventive care is one side of the quality of clinical care, being chronic disease management the other one. The indicators are: primary childhood immunisation, MMR (measles, mumps, and rubella), vaccine at 13 months, child preschool booster, preschool MMR booster, uptake of cervical cytology. To assess the quality of prevention care, for each practice a questionnaire was sent to the appropriate health authority to collect information on rates of uptake for cervical cytology screening; primary childhood immunisation; measles, mumps, and rubella immunisation; and preschool vaccination.

[19] The indicators applied are: team climate, continuity of care, receptionist scale, interpersonal care, overall satisfaction. To assess the access and interpersonal care from a patient perspective, 200 adults from each practice list were randomly selected and a copy of the general practice assessment survey was sent to each patient. Patients in five out of the six health authority areas received two postal reminders. Data from these questionnaires were used to assess the quality of access, continuity of care, and interpersonal aspects of care. To assess team climate and team effectiveness, the team climate inventory was sent to all staff employed by the practices; 48 (80%) practices took part in this assessment. Practices where less than 30% of the staff completed questionnaires were excluded from the analyses. The analyses included data from 42 (70%) practices, representing 387 (60%) members of staff. The team climate inventory assesses perceptions of staff members of how people work together, how frequently they interact, whether teams have identified aims and objectives, and how much practical support and assistance are given towards new and improved ways of doing things. They combined the team climate subscales into a single score.

Advantages and disadvantages were outlined for both measures. What it could be concluded is that the choice of measure to assess quality is only one step. To apply a measure it needs to be 'confident that variations in measured performance accurately reflect variations in actual performance' (Goddard et al., 2002a – p. 509). Several are the sources of variations that can influence a performance indicator and are beyond the control of the GP (both systematic influences such as patients characteristics, institutional setting, external environmental and random influences which persist even after adjustment). It needs to control for them before infer about the performance of the physicians.

1.4 Models of physician behaviour and quality incentives. A framework

The economic literature on physician behaviour has developed several different theoretical models. Much of the literature has modelled physicians as self-employed individuals with their own objectives (utility) to maximise. Most theoretical models have included income (or net income) as the main argument. Others based on principal-agent theory also have included effort as an argument negatively related to utility. The rest of models have included an 'ethical' argument to take into account the professional codes of conduct; the patient's welfare; the patients' economic well being; and the interest of society. Many models have been also specified in the context of supplier induced demand. This strand of literature includes 'inducement' as an argument, assuming that doctors know that they are inducing demand for health care that is unnecessary. Several models have also suggested doctor's reputations and status, practice characteristics, intellectual satisfaction and autonomy as an argument in the utility function (see the remainder of this chapter for a detailed discussion of these issues). The vast majority of these studies have used treatment decisions as the main decision variable (i.e. referrals, prescribing, etc.).

To sum up, physician behaviour is not exclusively driven by financial incentives but also by status seeking, intrinsic benefit and altruism (Pauly, 980; Dionne and Contandriopoulos, 1985; Lerner and Claxton, 1994; Encinosa et al. 1997; Scott 2001); as well as the cost of the effort. Quality incentives can be attached to each

element of the utility function. The sources of these incentives are competition, regulation, the physician's ethical values, and social and professional norms (Kuhn, 2003). Tab. 1 contains a framework for considering GP's incentives. It combines the sources of incentives (rows) with the physician's objectives (columns).

Tab. 1 Classification of incentives (Souce: Khun, 2003 p. 42)

Physician objective **Source of incentive**	**Income**	**Status**	**Intrinsic benefit and altruism**
Competition	Demand response, quality competition	Status competition in income on 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	Reputation relative to professional or societal norm	Work ethic, internalised norms

Competition stimulates income related incentives if a physician's remuneration increases with the demand for his services and if this demand is reactive to quality. Quality elasticity of demand, in turn, relies on the patient's ability of measuring or experiencing the quality of a service and of obtaining the same or a substitute service from more than one provider.

Physicians may compete not only for income but also for social status. The closer is the reference group, the stronger is the status competition.

The impact of regulation on status as an incentive is ambiguous. On one side the publication of performance indicators or the implementation of peer review improve status competition by making an individual's performance common knowledge. On the other side, regulation may erode a social norm and thereby devaluate status as a source of motivation. For example, if good professional behaviour constitute 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

systems and other forms of income-related regulation are likely to play a more important role.
Social norms can play an indirect role in stimulating income related incentives by shaping how and to what extent competition and regulation act as sources of incentives.

Irrespective of the financial and social reward incentives, physician can be intrinsically motivated and they can derive a benefit directly from improving the health of their patients. However, external incentives may crowd out intrinsic motivation if a physician perceive the provision of quality as the result of competitive pressure or regulation rather than of his autonomy.

1.5 The patient-physician interaction

The economics of the primary health market is concerned with the interaction between third-payers, consumers and physicians. This section focus mainly on the patient-physician interaction and his implications on quality provision. The relationship between the third-party and physicians will be discussed with more details later on (see section on regulation). Both relationships are most frequently seen as ones of agency. In particular, agency relationship occurs between the patient and the physician when the former (the principal) delegates decision making authority to the latter (the agent). The health economics literature developed several approaches to model the physician–patient relationship moving from an old perspective, where the patient passively accepts doctors' advice and treatment, to an approach characterised by strategic interaction between patient and physician (Barigozzi and Levaggi, 2005).

The early approach depicts the physician as a perfect agent. A perfect agent is defined as one who makes the same decisions that the patient would have made if the patient possessed the same information as the agent (not necessarily perfect and complete information). Delegation to the physician is considered optimal and efficient since he is completely benevolent, perfectly informed about health matters and takes decisions for the patient only in his best interest (Feldstein 1970, Phelps 1992).

The old view shows a number of shortcomings, the most relevant being that it does not consider the physician's market power in influencing the consumer decision making (see the remainder of the chapter). The physician has significant informational advantages making opportunistic behaviours possible. His superior knowledge may concern the patient's health status (Arrow, 1963), the available treatments and their possible effects. Specifically, in primary care market, with the possible exception of routine and repeat services, doctors generally have more information and knowledge relevant to determining the medical conditions of their patients and the treatments that are most likely to be helpful. Furthermore, the physician's actions are often not observable by the patient and quality of services provided is sometimes not verifiable (among others Ma, 1994; Ma, 1997), even ex-post. Whenever his objective function is not perfectly aligned with

that of the patient, the physician can use this informational advantage to provide over treatment in order to increase the remuneration of his work (Evans 1974, Farley 1986, Mooney and Ryan 1993, DeJaegher and Jeger 2000 and 2001, McGuire 2000) or to provide low effort in order to decrease his disutility costs (among others Ma and McGuire 1997). The theoretical and empirical literature has extensively analysed the former problem (usually defined as 'demand inducement') which produces an increase in the price and volume of health care along with a decrease in the appropriateness of health care and patients' satisfaction. Improving information can help to reduce the opportunism of the physician. Also managed care can control agency by limiting physicians' discretion over healthcare quantities.

A more recent strand of the literature argues that the conjecture that more information is always preferred to less might not be valid when there are consequences of stress and anxiety associated with medical procedures or uncertain health outcomes (Barigozzi and Levaggi, 2005). This approach is known as the emotional agency approach and is motivated by a growing literature on behavioural medicine (Baum et al. 1997). Patients may 'rationally' postpone the resolution of uncertainty and deciding to stay ignorant even though such uncertainty implies taking less efficient actions (Koszegi 2003). The Psychological Expected Utility model proposed by Caplin and Leahy (2001) extends the standard model of choice under uncertainty in order to explain how anticipatory feelings such as anxiety or hopefulness influence decision makers. Utility depends not only on physical utility but also on the expectation of such physical utility. Since people derive utility directly from their beliefs, they must consider how the information they gather will affect those beliefs. It has been observed that attitude towards information is not uniform: some patients prefer to learn their health conditions, even if this means receiving bad news (early resolvers) while others, the more anxious, prefer to stay ignorant (late resolvers). Kozsegi (2003) shows that, if the patient is sufficiently anxious (i.e. information-averse), the choice of staying ignorant can be optimal even though it implies damage deriving from inefficient actions. Some recent studies show that physicians express difficulty in revealing sensitive information to their patients. In

particular, doctors take their patients' emotions into account and, more importantly, their recommendations and communication are affected (and even 'distorted') in response to them. In this respect, feelings add an important dimension of complexity to the doctor-patient interaction; two recent papers, Caplin and Leahy (2004) and Kozsegi (2005) investigate this issue. Caplin and Leahy (2004) analyse a show-and-tell game where the patient can be either an early or a late resolver patient and this information on his type is private. The patient decides whether to reveal his type to the physician or not. The physician observes what type of treatment the patient needs: the traetment can be characterised either by low or high risk. When the intervention is low risk the preferred outcome α is more likely than the worst one β, whereas when the operation is high risk the opposite holds. Once the message has been received by the patient (who can show or not show his type), the physician must decide whether to tell the patient which operation he needs. The physician is completely empathetic, that is he derives utility only from his beliefs in the patient's welfare.

The remainder of this chapter discusses three mechanisms through which physicians may influence the demand for medical services, namely by setting the quantity of a nonretradable service, by setting the level of a noncontractible input ('quality'), and in presence of asymmetric information, by persuasion (McGuire, 2000; Lien et al., 2004)[20]. The third mechanism is also known as supplier induced demand or moral hazard of the physician. An alternative theory to the supplier induced demand labelled as target income model is also reviewed and a synthesis of the two models is presented. Finally, we discuss some instruments for dealing with quantity and quality setting in the physician-patient interaction and how they can be used to reduce the opportunistic behaviour of the physician and to improve quality: competition, ethics and regulation (in the form of mixed payments and pay for performance).

[20] Observed utilization of medical care reflects both consumer and physician incentives (Zweifel and Manning, 2000). This thesis focuses on the second ones (i.e. on supply side).

1.5.1 The physician as a quantity setter of a nonretradable service

Standard economic theory assumes that in a competitive market there is a single homogeneous and *tradeable* commodity. By contrast, as Gaynor (1994, p. 224) observes in his review, 'services are by their nature inherently heterogeneous and nonretradable'. Accordingly, a particular medical service such as a diagnosis or a treatment provided to one patient cannot be resold by that patient to some other customer. The nonretradability of physicians services has important implications for price discrimination and more generally for price and quantity setting ((Farley, 1986; Gaynor, 1994; McGuire, 2000). Specifically, in a context of monopolistic competition and in absence of asymmetric information, McGuire (2000) show how an income profit-maximising physicians may influence quantity of care provided to patients by setting the quantity of a nonretradable service. This is known how the first mechanism through which physicians may influence patient decision making. We discuss the problem of quantity setting without and with price regulation and conclude with a third case which includes coinsurance from the patient side.

In McGuire's model (2000) the quantity of services that a physician supplies is x. A patient is assumed to benefit from the receipt of physician services according to the benefit function $B(x)$ which captures also time costs, inconvenience, health shocks, etc. The marginal benefit function is given by

$$b(x) = B'(x) \qquad [1.1]$$

where

$$b(0) > 0 \text{ and } b(x) > 0,\ b'(x) < 0 \qquad [1.2]$$

The negative second derivative occurs because of diminishing marginal utility of health status and because the health production function $H = H(x)$ has a negative second derivative.

Since providing care is costly, it is assumed that the provider faces a constant cost per unit c. If p is the price of physician services, the profit function is

$$\pi_{ph} = (p - c)x \qquad [1.3]$$

while the patient net benefit is

$$NB_p(x) = B(x) - px \qquad [1.4]$$

The efficient level of physician services x^* is given by the level which satisfies the following condition

$$b(x) = c \qquad [1.5]$$

Let $NB^*(x) = B(x^*) - c(x^*)$ the maximum possible patient benefit and x^m the level of x that maximize $B(x)$ or the solution to $b(x^m) = 0$. See Fig. 3.

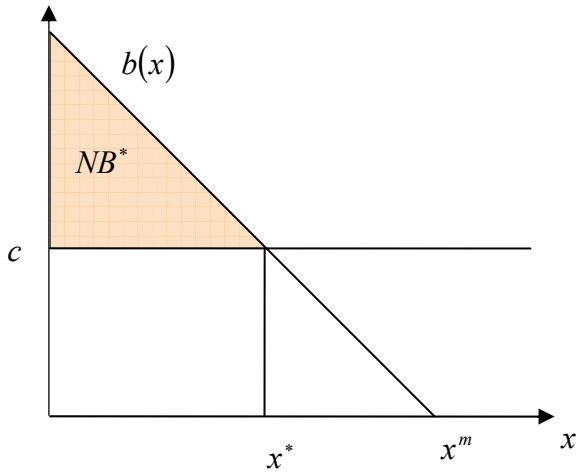


Fig. 3 Benefits and costs of physician benefits (Source: McGuire, 2000)

Since patients do not regard doctors as perfect substitutes - medical care being a personal service - let NB^0 the net benefit received by the patient if he leaves his physician and chooses another physician from whom to receive care. If $NB^0 = 0$ the patient has no alternative and the physician is a monopolist. If

$NB^0 = NB^*$ the market is perfectly competitive and the physician has not market power. In general $0 < NB^0 < NB^*$. The patient uses the current physician if and only if $NB \geq NB^0$. Given less than perfect substitutability, physician will supply services required to keep benefit patient he gets from his provision of services greater than next best option while maximizing income. The physician's profit maximization problem with complete information is described below. Two cases are distinguished: the former allows the physician to choose both price and quantity; the latter includes the effect of price regulation. The price and quantity of physician services are found by maximizing the physician's profit subject to the constraint on patient net benefit imposed by competition with alternative physicians.

Assume that both the price and the quantity can be chosen by the physician.

$$Max\ \pi_{ph} = (p - c)x \qquad [1.6]$$

$$\text{s.t. } B(x) - px = NB^0 \qquad [1.7]$$

The Lagrangian function is given by

$$\ell = px - cx - \lambda\left[B(x) - px - NB^0\right] \qquad [1.8]$$

The first order conditions are:

$$\begin{aligned} \frac{\partial \ell}{\partial p} &= x - \lambda x = 0 \\ \frac{\partial \ell}{\partial x} &= p - c - \lambda\left[b(x) - p\right] = 0 \\ \frac{\partial \ell}{\partial \lambda} &= B(x) - px - NB^0 = 0 \end{aligned} \qquad [1.9]$$

The price is determined so as to extract all surplus above NB_0

$$p = \frac{B(x^*) - NB^0}{x^*} \qquad [1.10]$$

Fig. 4 illustrates the solution. It is possible to notice the transfer of consumer surplus from patient to physician. The patient's lost net benefit of receiving care from this physician equals the loss of net benefits if the patient went to next best option. At price p the patient would prefer to consume less but nonretradability lets the physician set the quantity. It is clear that if $NB^0 = NB^*$ the physician is forced to deliver a quantity x^* at the competitive price, c. With exercising market power and with the nonretradability of services, the physician can implement a first-degree price discrimination. Patients with higher willingness to pay will pay more for the same services. Nonretradability shelters the price discrimination. The poor pay less because they have a lower willingness to pay not because they have a more elastic demand.

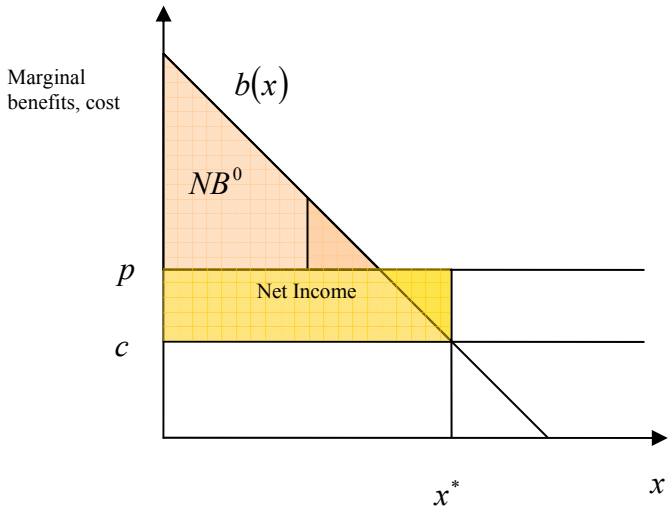


Fig. 4 Setting price and quantity with net benefit constraint (Source: McGuire, 2000)

Now we discuss the introduction of a price regulation. When payers set price, physicians still retain their market power by setting the quantity of their nonretradable service. The physician's profit maximization problem with complete information and price regulated by the third-payer is described below.

$$Max\ \pi_{ph} = px - cx \qquad [1.11]$$

$$\text{s.t. } B(x) - px = NB^0 \qquad [1.12]$$

The Lagrangian function is

$$\ell = px - cx - \lambda\left[B(x) - px - NB^0\right] \qquad [1.13]$$

The first-order conditions are

$$\begin{aligned} \frac{\partial \ell}{\partial x} &= p - c - \lambda\left[b(x) - p\right] = 0 \\ \frac{\partial \ell}{\partial \lambda} &= B(x) - px - NB^0 = 0 \end{aligned} \qquad [1.14]$$

The net benefit constraint can be solved for quantity x. Note that the net benefit constraint implies

$$\frac{dx}{dp} = \frac{-x}{p - b(x)} < 0 \qquad [1.15]$$

i.e. a decrease in price will be cause an increase in quantity. Since the physician can not extract surplus by setting price he will extract surplus by setting quantity at a higher level. The fact that the price is greater than the cost ensures to reach this objective.

Now, assume that the price paid by the patient is less than the price received by the physician. Let θ be the coinsurance rate with $0 < \theta < 1$ and $p > c$ to ensure physician participation. Physicians decide only x. The maximization problem becomes:

$$Max\ \pi_{ph} = (p - c)x \qquad [1.16]$$

$$s.t.\ B(x) - \theta px = NB^0 \qquad [1.17]$$

The Langrangian becomes

$$\ell = px - cx - \lambda\left[B(x) - \theta px - NB^0\right] \qquad [1.18]$$

The first-order conditions are

$$\frac{\partial \ell}{\partial x} = p - c - \lambda\left[b(x) - \theta p\right] = 0 \qquad [1.19]$$

$$\frac{\partial l}{\partial \lambda} = B(x) - \theta px - NB^0 = 0 \qquad [1.20]$$

The physicians choose to provide the level of services given by

$$x' = \frac{B(x') - NB^0}{\theta p} \qquad [1.21]$$

Then, when price is constrained the doctor exercises market power by setting quantity beyond the point the patient would choose given the price he faces (see Fig. **5**).

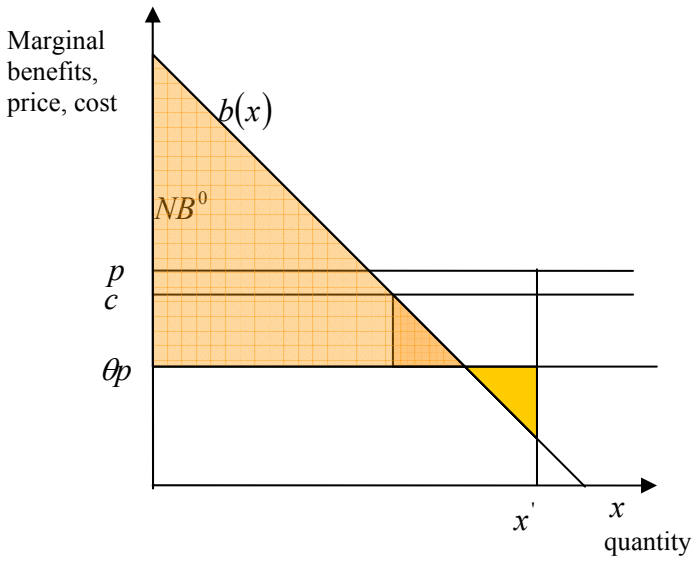


Fig. 5 Setting quantity with administered prices and insurance (Source: Mc Guire, 2000)

1.5.2 Choice of noncontractable input

The second mechanism concerns with a noncontractible input supplied by the physician in the health production process. This noncontractible input can be regarded as quality and can be referred to as the time and effort spent by the physician in treating the patient. This is additional to other measurable inputs such days, visits, or tests. Since effort is costly, the physician faces a cost function increasing in the level of effort. The marginal cost function increases as well, i.e.

$$c_e > 0, \qquad \text{and} \qquad c_e^{'} > 0$$

Quality is a productive input which affects health outcomes as well as the patient's demand. Consequently, the net benefit function depends on both inputs

$$NB = B(x,e) - p_d x \qquad B_x > 0;\ B_{xx} < 0;\ B_e > 0;\ B_{ee} < 0 \qquad [1.22]$$

where the price paid by the patient for each unit p_d is set by the payer. The number of patients the physician serves becomes a positive function of the net benefit $n(NB)$.

Since quantity is contractible while quality is only observed by the patient and it cannot be paid upon by a payer since it is unverifiable, a payer can only mitigate the incentive to over or under provide quantity and/or effort by introducing a per-patient contract with supply-side cost sharing as

$$R + p_s x \quad R > 0 \qquad 0 \le p_s < c \qquad [1.23]$$

where R is the capitation amount made independent of the services provided while p_s is the payment per unit of service which is less than c (see section on regulation for a detailed discussion on supply cost sharing and policy implications).

The profit maximization problem becomes

$$Max\ \pi = n(NB)[R + (p_s - c(e))x] \quad [1.24]$$

$$s.t.\ B(x,e) - p_d x = NB \quad [1.25]$$

The first order conditions are

$$n'(B_x - p_d)[R + (p_s - c)x] + n(p_s - c) = 0 \text{ with respect to quantity} \quad [1.26]$$

$$n' B_e [R + [p_s - c]x] - nc_e x = 0 \text{ with respect to quality} \quad [1.27]$$

These can be rewritten as

$$\frac{B_x - p_d}{NB/x}\left[\frac{R/x + p_s - c}{p_s - c}\right] = -\frac{1}{\varepsilon_{n,NB}} \text{ where } \varepsilon_{n,NB} = n' \cdot \frac{NB}{n} \quad [1][1.28]$$

$$\frac{R/x + p_s - c}{c} = \frac{\varepsilon_{c,e}}{\varepsilon_{n,e}} \text{ where } \varepsilon_{c,e} = c' \frac{e}{c},\ \varepsilon_{n,e} = n' \frac{\partial NB}{\partial e} \cdot \frac{e}{n} \quad [2][1.29]$$

From [1] the physician chooses to set quantity to satisfy

$$\frac{B_x - p_d}{NB/x} = -\frac{1}{\varepsilon_{n,NB}} \text{ when } R = 0 \quad [1.30]$$

Physicians have the incentive to set quantity beyond the point the patient would demand given his insurance ($B_x = p_d$)[21]. However, demand response ($\varepsilon_{n,NB}$) may limit the physician in setting quantity. When this elasticity is very high, as in a competitive environment, the net marginal benefit ($B_x - p_d$) tends to be zero.

When $R > 0$, the payment-system term matters and since $p_s - c$ is negative, the sign of the marginal benefit is reversed. The physician tend to limit quantity within what the patient would demand. In general, by fixing a high capitation

[21] Up to the average net benefit NB/x, the marginal net benefit $B_x - p_d$ is negative.

amount and reducing the payment per unit of service the payer can introduce an incentive to limit the over production of services.

As for the quality, the [2] implies that as the effort increases, the average net revenue per unit of service ($R/x + p_s - c$) increases as well. Since the effort is not contractible, what matters is the average profitability of services. As effort increases, the (positive) right-hand side of [2] rises being the marginal costs increasing in the level of effort and the marginal benefit falling.

1.5.3 The physician-induced-demand (PID)

Supplier-induce-demand occurs when physician does not act as a perfect agent for his patient because of financial incentives for providing extra services. In other terms, physicians do not act in the patient's best interest but they 'persuade' him to consume more care (Fuchs, 1978; Pauly, 1980; Eisenberg, 1986; Culyer, 1989; Williams, 1998 and many others). 'Persuasion' is regarded as an unproductive input. Positive definitions focus primarily instead on the physician's ability to 'shift a patient's demand curve to the right' (Richardson, 1981). But only rarely analysts explore the critical issue of how the extra services affect the patient's health (Pauly, 1979; Rosen, 1989). In addition, Dranove (1988) develops a framework in which inducement is limited by the patient's suspicion of the provider's aggressiveness; other approaches include the resource cost of the activity (Stano, 1987), and increasing professional discomfort (Evans, 1974; McGuire and Pauly, 1991). More relevant are the questions concerned with (Hadley et al. 1979) the kind of additional services provided when utilization rate increases; the access for the poor, disabled and aged; the type of additional diagnostic services and consultations in addition to the already well served and, finally, the overuse of technology-intensive services, with high real resource costs but low time costs to the user.

An extended conceptual framework for examining PID is presented in Tab. 2. It offers an analytically distinction among different types of utilisation that have been implicit in debates about inducement. The two defining dimensions of the framework capture the primary concerns in these debates and are related with two positive questions $a)$ would the patient have demanded the service if he had the

same information as the physician? (perfect *vs* imperfect agency) *b*) Did the services contribute positively to the patient's health status? A six cell matrix is defined.

Tab. 2 Physician-initiated utilisation (Source: Labelle et al. 1994)

		Effectiveness of Services		
		Yes	No	
			Neutral	Detrimental
Effectiveness of Agency	Yes	I	III a	III b
	No	II	IV a	IV b

The table classifies physician-initiated utilization along two dimensions: *a*) the effectiveness of the agency relationship (which has been a primary concern in the literature); *b*) the effectiveness of services utilized. The focus of investigation is if the physicians act in the best interest of the patient and if the patient is better off as a result of the utilization. The first cell of the matrix, containing type 1 utilisation, avoids many of the negative connotation typically associated with PID. Health-care utilisation is good in that it would have been demanded by the patient if he had full information, and it contributes positively to health status. Utilisation in cell *II* also improves health status, but there is disagreement between the physician and the fully informed patient as to whether the service should be consumed. Utilisation in this cell may occur for several reasons. The first one is that the physician' evaluation of the health outcome itself differs from the patient's; the second one is that physician may feel professionally compelled to provide for the patient all services known to contribute positively to health, regardless of the (small or zero) impact of the health improvement on the patient's utility; third, the physician may be acting out of financial self-interest; finally, in the case of externalities arising from non-consumption of services, the physician may in fact be acting in the social interest (including that of the patient's family for instance) rather than in the interest of the individual patient. The challenge here is to make patient's preferences more explicit to physicians. Type *IIIa* and *IIIb* utilizations are distinguished by the effect that each has on health status. While type *IIIa* neither improves nor worsens the patient's condition, type

IIIb has a detrimental effect on health. In both cases, however the physician is acting as a perfect agent in advising the patient to use the services. Utilisation in these two cells can therefore be attributed to ignorance or uncertainty on the part of the physician and the patient. Much of the work by Wennberg et al. (1982) suggests that this type of utilization exists especially for discretionary services such as tonsillectomies and colonoscopies. Categories *IVa* and *IVb* refer to the type of PID that has traditionally had negative connotations- utilization that is ineffective (*IVa*) or harmful (*IVb*) to patients and which informed patients would not have chosen to consume. Both utilisations are undesirable because, like type *III* , they carry with them unnecessary opportunity costs.

1.5.3.1 Policy implications of the PID

The matrix in table 3 can be used as framework to derive policy implications which typically are discussed in the context of cost containment. In a system characterized by private or public third payer, effectiveness of services is likely to be more important than is effectiveness of agency. In order to discuss allocative efficiency, it is necessary that the level of analysis be shifted from individual to the population. Other policy concerns which are not discussed here include the impact on health status, equity and distributional implications and the net social benefit, if any, of PID.

First, the utilization of services that do not contribute positively to health status and the fact that full informed patients would not have consumed constitutes a loss of welfare (*Cell IV*). Reforms of the reimbursement system (supplemented by legal recourse for malpractice in cell *IVb*) is the usual policy prescription. The reform may involve changing from fee-for-services to other payment methods such as salary, capitation and prospective payment per case which tend to remove the incentives to perform additional services. It may be possible to preserve fee-for-services payment but to monitor and regulate physician behaviour through mechanisms such as expenditure target, caps, direct negotiation of utilization patterns with physicians as a condition of payment (as in Canada and USA). By supplementing payment reform with other actions (information dissemination)

yield a reduction in the induced services as well, such as make information about effectiveness more available to patients.

Cell III implies a different problem. Here completely informed patients would agree with their physicians about the benefits of utilization but both are wrong. The source of the induced utilisation is not imperfect agency but lack of knowledge concerning effectiveness on the part of both physicians and the patient. If the lack of knowledge is due to the absence of evaluative studies of the service, one policy response is to support the conduct of such studies (in this case they cannot be categorised using the matrix). In case information exists about the effectiveness of the service the first policy is to disseminate this information to physicians (but this not necessarily guarantees a change in behaviour); the production of practice guidelines; the use of informal communication channels and local opinion leaders; continuing medical education. A complete policy response may need to include the creation of environments (such as those in competitive markets involving one or more Health Maintenance Organisations) in which consumers can benefit from choosing efficient, effective providers as source of care.

We have showed how interaction between patient and physician determines the quantities of care supplied to patients. Tab. 3, adapted from McGuire (2000), summarizes the three mechanisms in the literature. Common to almost all papers on this topic is a monopolistically competitive market structure, but different informational and contractibility assumptions have been made.

Tab. 3 Determinants of provider–patient interactions on quantity of health care (Source McGuire, 2000)

	Nonretradability allows quantity setting	Choice of noncontractible input	Persuasion
Market structure	Monopolistic competition	Monopolistic competition	Monopolistic competition
Information	Complete	Complete	Asymmetric
Physician's action in influencing use	NA	Not contractable	Unobservable
Main features	Supply determination within demand constraints; can explain inverse (P,Q)relationship	Demand response to 'quality' or some other physician input	Physician takes actions to persuade; constrained by demand response or ethics
Illustrative paper	Farley (1986)	Ma and McGuire (1997)	Dranove (1988)

1.5.4 The target income model (TI)

The target income hypothesis suggests that physicians use their informational advantage in order to achieve a target level of income or an improvement in their relative income. McGuire (2000) considers a simplified form of the target income model where the physician's utility depends only on net income, $\pi = (p_s - c)x$, and demand inducement, I. Physician chooses x by inducing demand in order to achieve a target, T, i.e.

$$x = \frac{T}{m} \qquad [1.31]$$

where the margin m is the difference $p_s - c$.

To formalize the physician behaviour under the target income hypothesis, let assume I be the number of induced unit of service. If these induced units are sold at a constant profit rate of $m = p_s - c$, and if the maximum profit level with zero inducement is in Q_0, then the physician's total profit (income) can be described by

$$\pi = mQ_0 + mI \qquad [1.32]$$

This linear equation represent the attainable combinations of π and I available to the physician, and it acts as the physician's budget constraint.

The indifference curves in Fig. 6 represent the physician's preferences. They slope upwards because inducement is a bad good. In order to stay on the same indifference curve, the physician must gain compensation in extra profits to offset the disutility of engaging in a higher level of I. Higher curves are preferred.

The figure can now be used to describe PID behaviour. The physician seeks to reach the highest indifference curve attainable, i.e. the point of tangency E_i. When competition increases, i.e. the average profit rate reduces from m to m', a new budget constraint with a flatter slope and a lower intercept is obtained. The

new equilibrium occurs at point E_2. In this example, increased competition has led the physician to choose a higher degree of inducement $(I_{E2} > I_{E1})$.

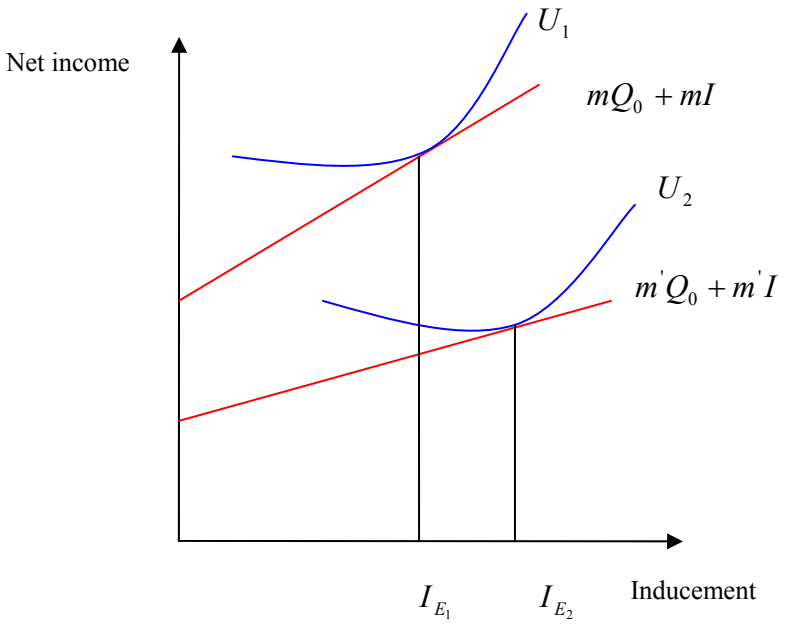


Fig. 6 Target income behaviour in a disutility of discretion model

1.5.4.1 A synthesis of the two models

The blue line in model *a* describes the physician's changed inducement in a different way (Fig. **7**). It removes income from the physician until he attains the equivalent utility as under the shifted m; thus it translates the drop in utility into an equivalent drop in income. The change in inducement that occurs by removing income is called the 'income effect'. In model *b*, the income effect is zero (the equilibrium inducement does not change when income is 'removed', i.e. $E_1 = E_2'$). Note that in model *b* physician would reduce inducement when faced with greater competition and a lower m. The new equilibrium is at E_2. The McGuire-Pauly (1991) synthesis tells us the size of the income effect is critical to identify and understand PID behaviour. A lowered profit rate m has two offsetting effects on inducement: *i)* substitution effect if inducement is less

profitable (smaller m), providers would do less inducement, or substitute away from it; ii) income effect, decreased income would make inducement more desirable. The income effect due to a lowered m must be positive and large enough to compensate a negative substitution effect.

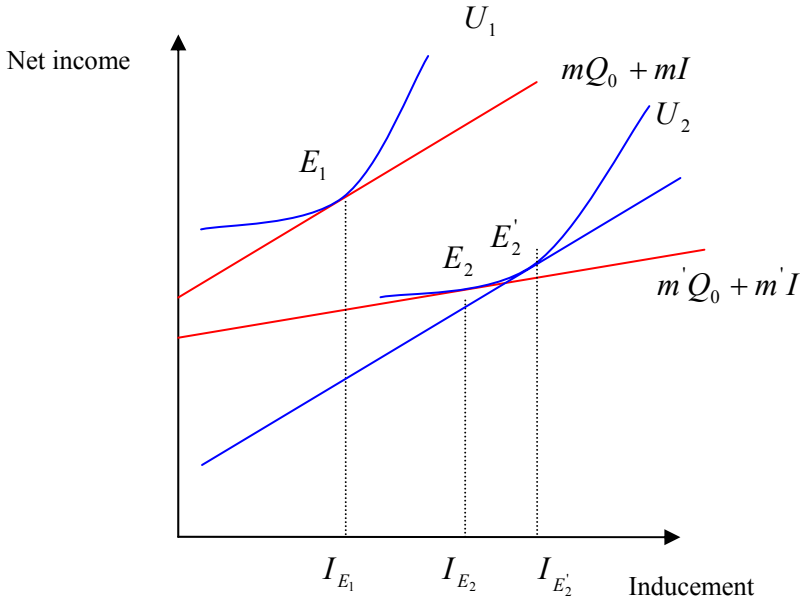


Fig. 7 *a* Target income behaviour (Source: McGuire and Pauly,1991)

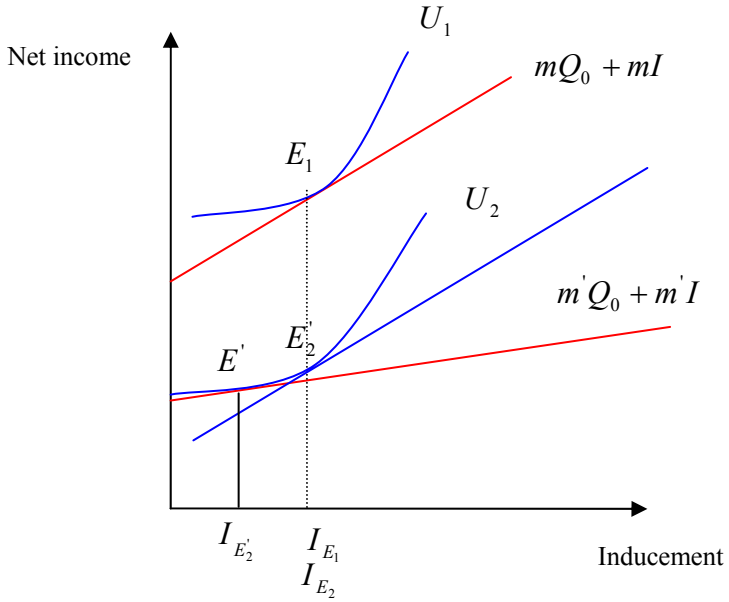


Fig. 7 *b* Profit-maximising behaviour (Source: Source: McGuire and Pauly,1991)

1.6 Small area variation in physician practice style

Small area variation (SAV) refers to the wide inter area variation in the per-capita use rates found for many medical and surgical procedures. The hypothesis underlying this phenomenon is that even if the physician acts as a perfect agent, differences in treatment patters may still occur due his uncertainty and ignorance over the best medical practice[22]. This is also known as the practice style hypothesis model. Under this hypothesis, it is important to investigate whether the observed variations are evidence of unnecessary or inappropriate care. Solving this controversy it is crucial in terms of policy. From an economic perspective, an adequate diffusion of medical information could help to achieve substantial welfare gains for society. The model assume that the practice style is determined by the physician's belief about the true production function for the treatment of a given health condition (which is unknown to the physician). He must choose the diagnosis and the treatment based on his available information and belief s. Fig. 8 shows the true production function (S^*) and the production functions of two individuals, one less optimistic (S_1) then the other one (S_2) about the effect of medical care on the patient's health care. The latter will be more likely to recommend the service than the former. If physician of type 1 and type 2 are not distributed similarly across markets then the medical care will vary accordingly.

This model also shows why it may be difficult to distinguish PID from SAV. Areas with high proportion of doctors like type 2 will have high level of utilization rates tempting some observers to conclude that PID is prevalent, whereas the true source is practice style. In alternative, physicians in those areas are simply acting as good agents by supporting their patients beliefs and preferences. In this case, the variations are due to patient preferences rather than SID or SAV.

[22] Source of variation other then practice style and clinical uncertainty are: mobidity and demographic structure, randomness – in process and measurement -, artefacts – incomplete coverage, or different terminology-, supply and availability of resources (McPherson, 1990; Mooney, 1994).

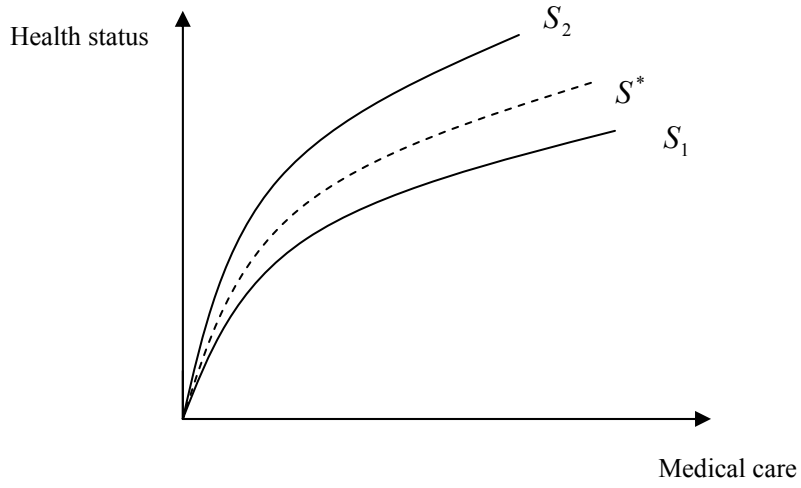


Fig. 8 Production function of a given health condition

The favourite measures of SAV are the coefficient of variation CV and the systematic component of variation (SCV). The CV is the ratio of a measure of the spread of on observed medical use rates across the small areas in the study region to the mean of the same measure. The measure of the spread is the standard deviation. Dividing this by the mean, the researcher adjusts for size of the rate being used. To compute the SCV the researcher first expunges from the observed treatment rate that portion of its variation that can naturally. The remaining portion might then become a better measure of the systematic effects of factors such as physician practice habits, and supply and demand factors.

A measure of the welfare loss due to SAV is proposed by Phelps and Parente (1990), and Phelps and Mooney (1993). The measure is based on the coefficient of variation for residual from regression of utilization rate on socioeconomic factors CV and on the absolute value of the price elasticity of demand E_p.

$$W = \frac{0.5(Total\ spending\ on\ X)CV^2}{E_p} \qquad [1.33]$$

Phelps (1995) applies the concept of consumer's surplus to explain the losses due to variations in utilisation rates (

Fig. **9**). The true marginal benefit (MB) curve for medical care involving a treatment R is given by the line MB^*. It reflects the benefit as perceived by the patient with full and complete scientific knowledge.

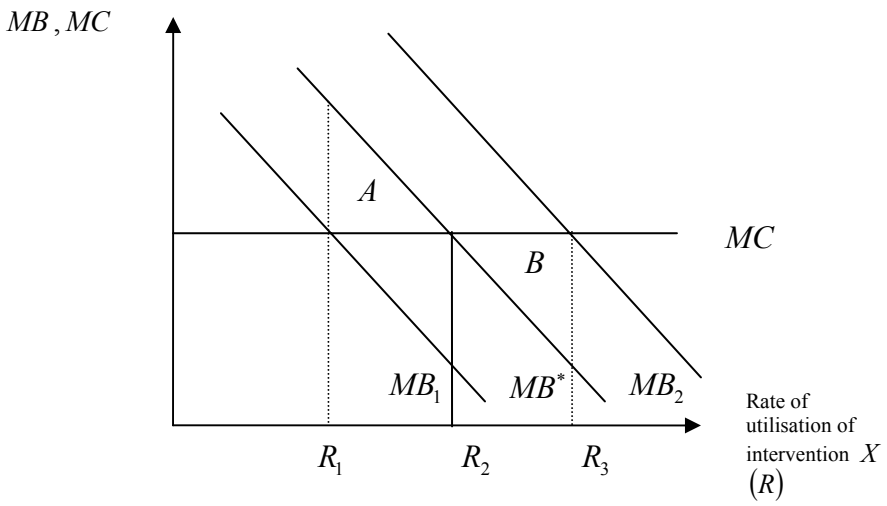


Fig. 9 The cost of inefficient variations

1.7 Competition as a source of quality provision

The proper role of competition in health care markets has long been debated. In general, physicians may compete on price and/or on quality but, in practice, since in several countries the health care markets are highly regulated, the non-price competition is more likely to rise. In this situation, most of physicians will engage in non-price competition in an effort to maintain and\ or increase total revenue. The impact of competition will depend mainly on the responsiveness of the consumer of health care to both quality and price and on the interaction between the nature of the market and the method of price setting. As for the first of these aspects, when quality is observed accurately but price is observed poorly demand becomes less responsive to price, allowing physicians to raise their prices but also giving them an incentive to increase and possibly over provide quality. Conversely, if price is observed accurately but quality is observed poorly, then the levels of quality supplied will be too low. Finally, if quality has several attributes, one of which is easier to observe than another (for example, clinical quality and patient amenity), then competition may lead to over provision of the one that is easily observed and under provision of the one that is less easy to observe (Chalkley and Malcomson, 1998, 2000; Dranove and Satterthwaite, 2000). As for the interaction between the nature of the market and the method of price setting, the form of this interaction will affect the general level of quality. In a market in which consumers of health care are covered by generous health insurance (as in the United States before the 1980s), they will not be sensitive to price, but will be responsive to differences in quality. So price may be high, but quality can also be high. In markets where consumers have 'harder' budget constraints (as in the UK during the 1990s internal market), price may be more important and physicians will compete on prices, leaving quality to fall below efficient levels. Finally, where a single price is fixed for all providers for a treatment (as in the current arrangements in England), there will also be no price competition. In this case, all competition will be in terms of quality. Competition may lead to excessive levels of quality and excessive product differentiation (Gaynor 2004). But if government reimbursement for a treatment is too low, competition may lead to the quality of this treatment being too low.

Despite the distinctiveness of these regimes, the competition among physicians to which each gives rise can be understood through a common model: monopolistic competition. A monopolistically competitive market consists of product differentiation and a sufficient number of physicians each of them maximise their own objective given the actions of their competitors. Each physician has some market power who faces an imperfectly elastic demand function for his service, can set a price above marginal cost, or choose his service level such that marginal cost falls short of a fixed fee for service, such that he can make a short-run operating profit.

1.7.1 Model 1: Price and quality competition

In this section we discuss in more details how competition may be enhanced through the demand response mechanism. We assume that the physician may set both price and quality of services provided. This means that demand for physician's service depends both on price and on quality. According to Chalkley and Malcomson (1998, 2000) and Satterthwaite (2000) patients are usually able to observe some quality attributes better than others. For example, they can be good judges of the amenities and friendliness and bad judges of clinical quality or medical skills. From a physician's perspective, there is then a specific elasticity of demand attached to each aspect of quality and the elasticity with respect to amenities to be likely greater then the elasticity with respect to clinical quality or medical expertise. The demand function can be written as

$$q\left(\overset{-}{p}\overset{+}{x},\overset{+}{y},\overset{+}{\hat{p}}\overset{-}{\hat{x}},\overset{-}{\hat{y}},\overset{-}{n},Z\right) \qquad [1.34]$$

Demand decreases in own price p, increases in own clinical quality x and patient amenities y, increases in competitors' clinical quality $\hat{x}$ and amenities $\hat{y}$, decreases in number of providers n and depends on purchaser attributes Z (e.g. type – if patient or institution -, numbers, information, degree of purchaser cost sharing).

The physician chooses the level of each dimension of quality and this is likely to results in a bias towards those dimensions that are easy to observe and to which patient demand is more sensible.

Assume the variable cost from providing care is of the form

$$C(q,x,y)=(a+bx+cy)q\,;\ a,b,c>0 \qquad [1.35]$$

and increases in quality and amenities.

In this circumstance, doctors seek to maximise their income which is given by the difference between the revenue and the variable monetary cost of providing care

$$\pi = pq(\cdot)-C[q(\cdot),x,y]=(p-a-bx-cy)q(\cdot) \qquad [1.36]$$

Each provider maximises profit by choosing x and y, while taking as given the competitors' choices of $\hat{x}$ and $\hat{y}$ (Nash behaviour). No provider wishes to deviate from his profit maximising choice p^*, x^* and y^*, given the competitors' choices $\hat{p}^*$, $\hat{x}^*$ and $\hat{y}^*$. Formally:

$$\max_{x,y}\pi(p,x,y,\hat{x},\hat{y},n,Z) \qquad [1.37]$$

The first order conditions are

$$\frac{\partial\pi}{\partial p}=q(\cdot)+(p-a-bx-cy)\frac{\partial q}{\partial p}=0 \qquad [1.38]$$

$$\frac{\partial\pi}{\partial x}=-bq(\cdot)+(p-a-bx-cy)\frac{\partial q}{\partial x}=0 \qquad [1.39]$$

$$\frac{\partial\pi}{\partial y}=-cq(\cdot)+(p-a-bx-cy)\frac{\partial q}{\partial y}=0 \qquad [1.40]$$

Reformulating the first-order conditions in terms of the elasticities with respect to price η_p, clinical quality η_x and amenities η_y

$$\eta_p = \frac{\partial q}{\partial p}\frac{p}{q(\cdot)} < 0 \qquad [1.41]$$

$$\eta_x = \frac{\partial q}{\partial x}\frac{x}{q(\cdot)} > 0 \qquad [1.42]$$

$$\eta_y = \frac{\partial q}{\partial y}\frac{y}{q(\cdot)} > 0 \qquad [1.43]$$

and solving for the optimal values p^*, x^* and y^*

$$p^* = \frac{\eta_p}{1+\eta_p}(a+bx+cy) = \frac{\overbrace{a\eta_p}^{<0}}{\underbrace{1+\eta_p+\eta_x+\eta_y}_{<0}} \qquad [1.44]$$

$$x^* = \frac{\eta_x}{1+\eta_x}\frac{\overline{p}-a-cy}{b} = \frac{(\overline{p}-a)\eta_x}{b(1+\eta_x+\eta_y)} \qquad [1.45]$$

$$y^* = \frac{\eta_y}{1+\eta_y}\frac{\overline{p}-a-cy}{c} = \frac{(\overline{p}-a)\eta_y}{c(1+\eta_x+\eta_y)} \qquad [1.46]$$

Note that $\eta_p < -(1+\eta_x+\eta_y)$ is a necessary condition for the existence of meaningful equilibrium. Quality elasticities must be sufficiently low relative to price elasticities.

From first equality clinical quality and amenities are complements

$$\frac{dx^*}{dp} > 0\,;\ \frac{dy^*}{dp} > 0\,;\ \frac{dp^*}{dx} > 0\,;\ \frac{dp^*}{dy} > 0 \qquad [1.47]$$

$$\frac{dx^*}{dy} = \left(\frac{dx^*}{dp}\right)\left(\frac{dp^*}{dy}\right) > 0 \text{ and } \frac{dy^*}{dx} = \left(\frac{dy^*}{dp}\right)\left(\frac{dp^*}{dx}\right) > 0 \qquad [1.48]$$

and the optimal level of each variable increases in the levels of the other two variables.

The reaction function diagrams are illustrated in figure 8. The first refer to price-quality relationship while the second one to the amenities-quality relationship. When the price increases, the marginal profit on each unit of service increases as well and, as a consequence, the incentive to stimulate demand by quality and amenities arises. When clinical quality improves, this leads to set a higher price. Since profit margin with respect to clinical quality increases, the physician will have an incentive to stimulate demand by amenities.

From the second equation it is evident that a reduction in the elasticity of the demand is followed by a higher level for each strategic variable. Conversely, a more elastic demand with respect to both aspects of quality induces a lower level in price and quality provision (see Fig. 10).

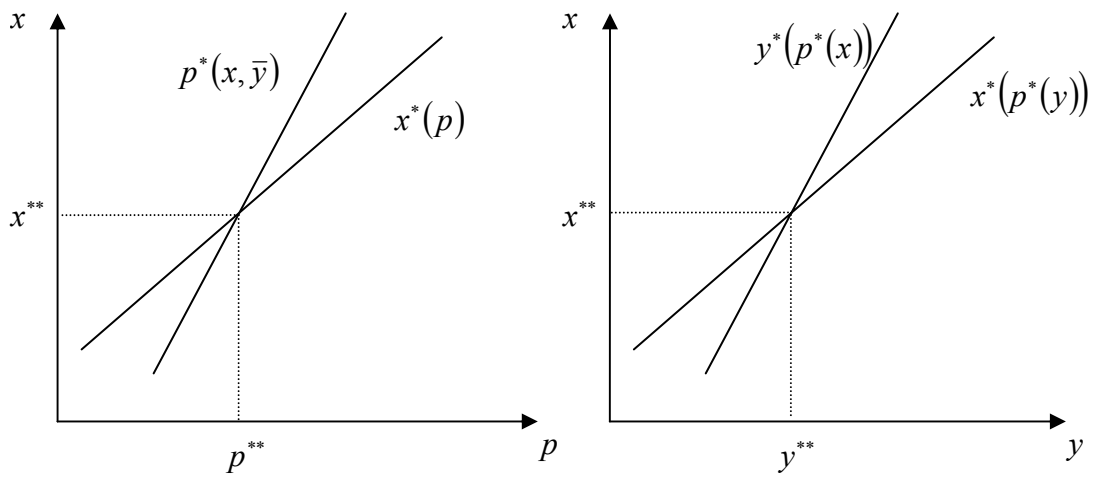


Fig. 10 Reaction function diagram $a)$ Price quality $b)$ Amenities-quality

It is possible to draw some predictions from the previous model. First, if patients are insulated by insurance, it is likely that the elasticity with respect to price is low (low $|\eta_p|$) and this will lead to an over-provision of quality and amenities at an excessive price. If the patients have incomplete information on clinical quality

which reflects in a low elasticity with respect to clinical quality, then the physician will under provide clinical quality relative to amenities. In this case, performance indicators may improve information on quality. This two case are embedded in the so called *patient driven competition*. Second, when primary care physicians act as patient's agents and are paid by a third-payer, a higher elasticity with respect to clinical quality (high η_x) will induce physicians to over provide clinical quality. This case is also known as *physician driven competition.* Third, in presence of price-conscious purchasers such as health plans or budget constrained purchasers, a high elasticity with respect to price will lead to both lower price and quality levels, if the latter is not controlled by audit. This gives rise to *payer driven competition*. Fourth, provided that the number of physicians raises all elasticities, the effect on competition depends on relative increases. Specifically, when

$d|\eta_p| < d\eta_x$ quality competition is enhanced (US health care reform during 80s)

$d|\eta_p| > d\eta_x$ price competition is enhanced (80s reforms: prospective reimbursement, managed care)

1.7.2 Model 2: Quality competition under fee regulation

In this section we discuss in more details how competition may be enhanced through the demand response mechanism when physician receives a fee for their services which is fixed centrally by the government ($\overline{p}$) and patients are insulated by insurance. This means that the price of a medical service is perceived to be zero by consumers and demand depends on quality and amenities alone. Consequently, even thought the price is zero the demand for medical services can be still positively responsive to quality (Ma, 1994; Gravelle, 1999; Chalkley and Malcomson, 2000), where quality may be related to travel or time cost or to non-monetary disutility from receiving care, i.e. side effects or physical or psychological discomfort (Kuhn, 2003). The demand function is increasing in

both clinical quality and patient amenities and decreasing in both competitors' clinical quality and amenities as well as in number of providers.

$$q\left(\overset{+}{x},\overset{+}{y},\overset{-}{\hat{x}},\overset{-}{\hat{y}},\overset{-}{n},Z\right) \qquad [1.49]$$

Assume the variable cost from providing care is of the form

$$C(q,x,y)=(a+bx+cy)q \qquad [1.50]$$

In this circumstance, doctors seek to maximise their income which is given by the difference between the revenue and the variable monetary cost of providing care

$$\pi=\overline{p}q(\cdot)-C[q(\cdot),x,y]=(\overline{p}-a-bx-cy)q(\cdot) \qquad [1.51]$$

With a fee rate unchangeable, physician may increase income by improving quality to attract more patients provided that the demand is at some extent elastic. Each provider maximises profit by choosing x and y, while taking as given the competitors' choices of $\hat{x}$ and $\hat{y}$. Formally:

$$\max_{x,y}\pi(\overline{p},x,y,\hat{x},\hat{y},n,Z) \qquad [1.52]$$

The first order conditions are

$$\frac{\partial\pi}{\partial x}=-bq(\cdot)+(\overline{p}-a-bx-cy)\frac{\partial q}{\partial x}=0 \qquad [1.53]$$

$$\frac{\partial\pi}{\partial y}=-cq(\cdot)+(\overline{p}-a-bx-cy)\frac{\partial q}{\partial y}=0 \qquad [1.54]$$

Reformulating the first-order conditions in terms of the elasticities with respect to clinical quality η_x and amenities η_y

$$\eta_x = \frac{\partial q}{\partial x}\frac{x}{q(\cdot)} > 0 \qquad [1.55]$$

$$\eta_y = \frac{\partial q}{\partial y}\frac{y}{q(\cdot)} > 0 \qquad [1.56]$$

and solving for the optimal values x^* and y^*,

$$x^* = \frac{\eta_x}{1+\eta_x}\frac{\overline{p}-a-cy}{b} = \frac{(\overline{p}-a)\eta_x}{b(1+\eta_x+\eta_y)} \qquad [1.57]$$

$$y^* = \frac{\eta_y}{1+\eta_y}\frac{\overline{p}-a-cy}{c} = \frac{(\overline{p}-a)\eta_y}{c(1+\eta_x+\eta_y)} \qquad [1.58]$$

From first equality clinical quality and amenities are substitutes:

$$\frac{dx^*}{dy} < 0 \text{ and } \frac{dy^*}{dx} < 0 \qquad [1.59]$$

The reaction functions are illustrated in Fig. 11.

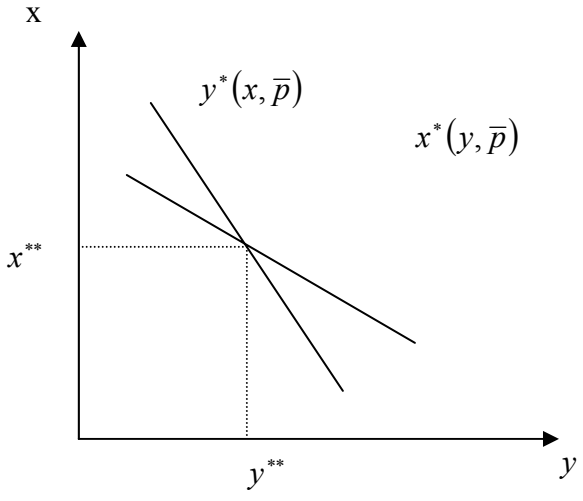


Fig. 11 Reaction function diagram

Compared with the price setting case, an increasing in the amenities leads to an increase in the marginal cost which reflects in a lower margin gained per patient

by delivering clinical quality. As a consequence, the level of clinical quality provided falls.

From the second equation some comparative static properties can be derived. Both clinical quality and amenities increase in fee $\overline{p}$. The higher is the fee per patient or treatment fixed by the payer/regulator, the greater the revenue derived from each unit of service provided is. Therefore, the physician is willing to supply to additional patients providing better quality to attract them. If the level of fee at which the payer can induce a certain level of quality is low than the quality elasticity of demand is great. This trade-off between the provision of quality and the financial feasibility becomes stronger as the elasticity reduces. In addition, clinical quality increases in own elasticity of demand η_x and decrease in elasticity of demand with respect to amenities, η_y. In general, the lower is the quality elasticity of demand the more reluctant the physician will be to provide quality for any level of fee. Provided that the demand is not inelastic to quality, a regulator/payer can stimulate the provision of quality by increasing the fee paid to the physician. More elastic is the demand the greater the provision of quality. If the elasticity is low a trade-off between providing quality and financial feasibility can arise. In this case it is important to understand the determinants of the quality elasticity of demand in order to implement the appropriate policy.

1.7.3 Quality elasticity of demand and its determinants

The main factors affecting the elasticity are summarised in Figure 10. These are the outside substitutes for the service (e.g. new technologies, pharmaceuticals, secondary care), the number of physicians or the physician density (i.e. the degree of competition), the information available to patients and the switching costs (Kuhn 2003).

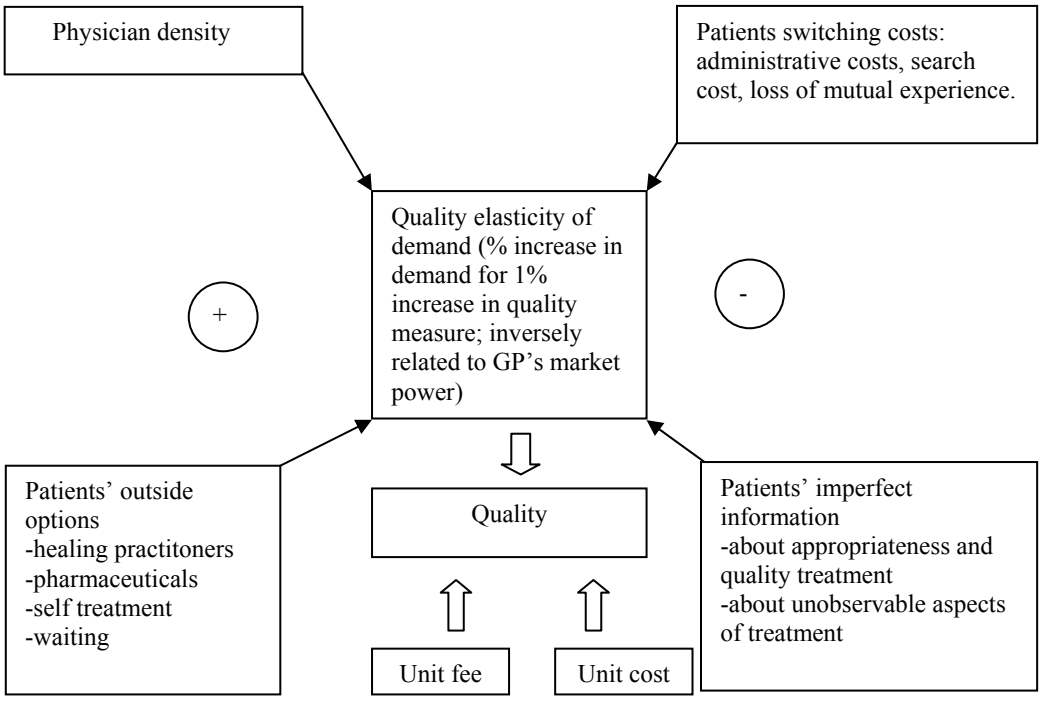


Fig. 12 Quality elasticity of demand and its determinants (Source: Kuhn, 2003)

1.7.3.1 Patient's outside options

The greater are the patients' outside options in terms of better substitutes for the primary care, the greater the quality elasticity of demand will be. This means that the physician exhibit a limited market power and the degree of competition is high. The increasing provision or promotion of services such as information or consultation via internet or telephone relating to self-treatment, the retailing of pharmaceuticals and health care equipment for domestic use, the supply of nursing or alternative medical services, imply an increasing of the outside competition. This can also be interpreted as a reaction of the physician to competition itself. In terms of policy, therefore, it can be recommended to improve quality by enhancing the availability of outside options to patients. For example, the English NHS promotes outside options by creating walk-in centres staffed by nurses than by physicians, access via telephone, internet, information points.

1.7.3.2 Patient's imperfect information and mechanisms to improve it

The level of information affects positively the quality elasticity of demand (Dranove and Satterthwaite 1992, 2000). When patients are weakly informed about quality they reduce the weight given to quality and become less responsive to quality changes. The incentive to provide quality for physicians is lower as well. In particular, they provide the higher effort when some dimensions of quality are easily observable by the patient than others. Asymmetric information about the physician's skills in providing medical care (hidden information) and effort (hidden action) restrict patient choice and reduce competition. Hidden information relates to 'structure' aspect of quality while the problem of hidden action relates to the process aspect of quality. If the patient is not able to distinguish between skilled and unskilled physicians, the skilled physicians attract the same demand and income as unskilled ones and cannot receive a return on their extra investments in skills. In this case, investments in skills will be too low and quality will be under-provided as effect of the 'adverse selection'. In terms of policy, this justifies medical accreditation procedures on the basis of a compulsory curriculum as a way of ensuring at least minimum investment in skills, continuing professional education programs, periodic revalidation of licences. In the other case, the patient does not observe effort directly (complete information) but he observes an imperfect indicator of effort, outcome. The patient is able to judge the outcome of the doctor's action, but on the other hand, outcome is affected by other unobservable factors (e.g. uncertainty). The patient does not know for sure whether the doctor's action was appropriate. In other terms, the patient cannot infer the effort provided by the physician in providing services of high quality. Physicians try to economise on monetary and non-monetary costs and therefore quality will be under-provided as effect of 'moral hazard'.

To restore these incentives, industrial organisation theory suggested a number of mechanism to reduce asymmetric information (Fig. 13). These solutions include patient search, signalling, reputation ad regulatory measures (kuhn, 2003).

Search costs are those which patients engage in to select their favourite physician. To reduce these costs it is recommended to provide public information on physicians for example through publishing performance indicators.

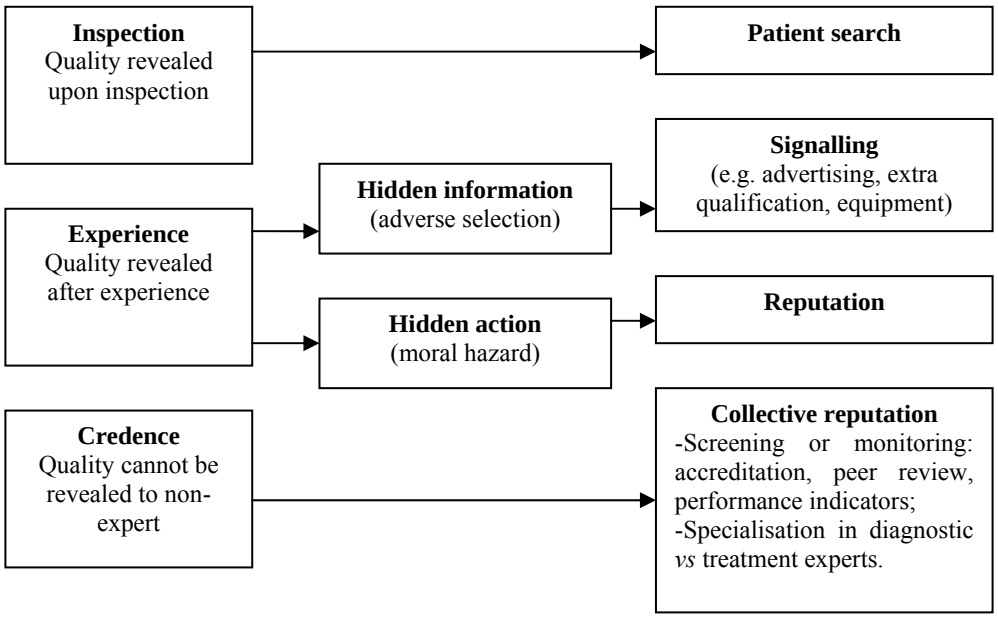


Fig. 13 Type of information and mechanisms of revelation

Signalling implies that the physician performs a costly activity observable by the patient and which only a skilled physician would engage in. The physician will choose a level of activity that is still profitable to him but not for the unskilled types. This type of activities may include advertising, extra services such as longer consultation hours or home visits; investments in medical equipment, continuity in medical education even over the level required. However, signalling is able to solve only the problem of hidden information. The problem of hidden action can be solved by the mechanism of reputation. Current experience of high effort does not implies a prediction of high effort in the future. As a rule, once a physician has established a reputation for providing quality he will continue to deliver good quality the greater is 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.

If quality is imperfectly observable even after experience it, the way to reveal information is by collective reputation which refers to the overall impression a patient receive of the profession or of a group of professionals. In the latter situation, physician's reputation is determined by the action of the other physicians belonging to the team (Getzen, 1984). Tirole (1996) points out that in an inter-temporal perspective, a bad reputation can be self-enforcing over time and in this circumstance only external interventions can restore quality incentives and professional's reputation.

1.7.3.3 Patient's switching costs

Patients face costs in switching to another doctor. These costs may include the time and trouble involved in changing registrations as well as a cost in terms of lower initial level of care ceteris paribus, the new physician being initially less well informed about the patient than his current doctor will. Medical records are an imperfect substitute for personal contact and are transferred with a significant delay (Gravelle and Masiero, 2000). When patients are involved in long-term relationship with a physicians, the latter may have an incentive to lower quality having the patient a disincentive to switch. Dranove and White (1987) stress the fact that a long term relationship reduces the patient's monitoring cost and the physician's cost of the diagnosis, where both kinds of costs are negatively linked to the quality of care. However, Gravelle and Masiero (2000) show that provision of quality is independent of the switching cost because of measurement error to switch[23]. By considering a model of quality competition between two physicians and across two time periods, they show how patients switch as the level of quality

[23] Giuffrida and Gravelle (2000) studied the implication of imperfect information and swithching costs for incentive for quality in a regulated market. The setting of the study was the UK primary health system after the reform introduced in 1990. This reform increased capitation fees and reduced patient' switching costs by abolishing the formal consent the current doctor had to give to the patient who wanted change the doctor. The objective was to stimulate quality by rising the potential profit a GP could gain by treating more patients. Since the fee is financed by taxation, patients care only about quality and location. The authors apply a standard Hotelling differentiation model. Patients make initial errors in judging quality and switching costs lock some of the mistaken patients into the wrong GP.

provided is lower than their expectations. This can be due to a patient's error. Since is more difficult to predict that a doctor is good than to predict that he is poor in delivering quality, it is possible to assume that the patient error increases with the level of quality. This implies that the associated cost of switching is an indirect cost of quality and the regulator chooses a level of quality under imperfect information that is below the optimum level under perfect information.

1.8 Regulation as a source of quality incentives

Regulation is usually employed as a stimulus of quality if competition is too weak or if it gives rise to dysfunctional incentives. Regulatory interventions include the design of payment system (soft regulation), the setting of quality standards, the introduction of audit systems, the publication of performance indictors, or the introduction of quality related performance pay (hard regulation).

Regulation gives rise to income related incentives by way of payment system. The mechanism of remuneration crucially determines in which way a physician's income is related to the quality of service. For example, under a capitation payment system a physician can use quality to attract patients. However, if patient's demand is not responsive to quality, capitation may induce the physician to reduce quality in order to save cost. Sometimes payments are directly linked to measures of quality, i.e. in the form 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 license.

1.8.1 Criteria for evaluating payment systems: micro-efficiency and macro-efficiency

Because of the trade-off between efficiency and quality which affects the health care production, provider incentive literature has become a relevant topic in health economics. The debate, in particular, is mainly based upon the evaluation of different payment systems, such as prospective and retrospective payment systems or a mix of both, designed by the payer in order to induce the provider to supply a desired behaviour. This section overviews this debate and then argues in favour of a mixed reimbursement as a mechanism able to mitigate the conflict between quality and efficiency.

This section is organized as follows. The first part provides a framework to compare purely retrospective and purely prospective payment systems, with respect to both micro-efficiency and macro-efficiency dimensions. The analysis is

then carried out by focusing on the advantages of the mixed payment system over the two types mentioned above. In the second part, it suggests which issues generally need to be taken into account when drawing up a doctor's reimbursement.

The analysis is intended to be normative. In other words, it considers what is appropriate for a government purchaser to do when he is concerned with social welfare.

Firstly, this section analyses the issue of efficiency with regard to both purely retrospective and prospective payment systems, focusing on their advantages and disadvantages. It then show the advantages of a mixed reimbursement system over the other systems.

Given the institutional context in which they exist, payment mechanisms may induce movement towards or away from efficiency, depending on the nature of the incentives. To determine the efficiency of the different payment systems, some criteria for evaluation need to be provided. Rochaix (1998) suggests considering two dimensions for efficiency, i.e. micro-efficiency and macro-efficiency, in order to evaluate different payment performances. Both of them involve the agency relationship. In particular, the physician is regarded as the agent of two principals. One principal is the patient, who is concerned to receive an appropriate treatment, and the other one is the payer (an insurer or a government agency), who is interested in the treatment that patients receive as well as keeping down the cost of providing treatments (Chalkley & Malcomson, 2000). Consequently, since the payer has a secondary aim which is cost-containment, a potential conflict of interest between the patient and the payer could arise. In the light of this tension, the concepts of micro-efficiency and macro-efficiency assume relevance. Micro-efficiency (in both allocative as well as productive meaning) does not involve actually any conflict between the payer and the patient, because the patient's preferences are taken into account. The type and the level of treatment are supposed to be very relevant and the use of factors be efficient for any given

treatment. In contrast, macro-efficiency refers to the physician-third-party payer agency relationship and it essentially refers to cost-containment. On the macro-side, what matters is the efficient overall level and structure of health care given the scarcity of funds. In this case, the optimal level of care depends on the opportunity cost of spending the same funds on other sectors such as education, retirement, etc.

Before evaluating the pure retrospective and prospective payment in terms of the efficiency criteria mentioned above, it is useful to clarify what each payment system mainly refers to. Broadly speaking, a pure retrospective payment (e.g. cost-reimbursement) concerns the case in which the provider is fully reimbursed for every item of healthcare, whereas a pure prospective payment (e.g. budget) regards the case in which the amount of reimbursement is predetermined for each condition. In this meaning, pure retrospective payment and pure prospective payment represent the extremes of supply cost-sharing (Ellis & McGuire, 1993). The former mainly has the advantages of enhancing freedom and greater continuity of care and the disadvantages of inducing demand for health provision and to provide expensive services. In contrast, the latter improves the coordination of care and keeps costs down, but leads to under provision of health care also in terms of insufficient effort and/or motivation. In addition, this type of payment has the incentive to select patients in order to attract the good prospects (cream-skimming), to discourage the more complicated or severe cases (skimping) or to do not provide care at all (dumping). It is clear that both retrospective and prospective payments are problematic from an efficiency point of view. In general, moving away from a retrospective payment towards a prospective payment system implies an incentive for providing a reduced quantity of services and *vice versa*. In particular, the retrospective payment system is likely to be macro-inefficient whereas the prospective is associated with micro-inefficiency. A way to overcome these disadvantages is obtained by introducing a mixed reimbursement[24].

[24] Several countries have adopted a mix of payment systems (Rochaix, 1998). Specifically, open-ended systems (such as Germany, France or Quebec) have moved towards prospective payment systems to solve macro-efficiency problems by introducing mainly negative incentives in the form of expenditure caps and targets. On the other hand, closed-ended

1.8.2 Prospective versus retrospective payments system: a formal framework

A formal framework is now introduced to show the advantages of a mixed payment over purely retrospective and prospective payment. Firstly, the utility function of a provider is analysed in order to illustrate the behaviour of the provider under different payment systems. Then, a supply curve of health care will be derived in order to discuss some policy implications.

In this model, the physician is assumed to receive a non-monetary benefit from treatment at level x. Non-monetary benefit may be related to either some characteristics the physician receives in terms of status, reputation or to altruism etc[25]. This benefit is given by $B(x)$, where x may concern the intensity of treatment, the number of patient visits to the doctors, the number of consultations, the effort of the doctor in terms of the time input, the observable dimension of quality, etc. The benefit increases in the level of care provided, while marginal benefits are decreasing, i.e.

$$B(x) > 0\,, \quad B'(x) \geq 0\,, \quad B''(x) \leq 0 \quad (concavity) \qquad [1.60]$$

Since providing care is costly, the physician faces a cost function increasing in the level of care (Fig. 14). The marginal cost function increases as well, i.e.

$$C(x) > 0\,, \quad C'(x) > 0\,, \quad C''(x) > 0 \quad (convexity) \qquad [1.61]$$

systems (such as US managed care) have moved towards mixed payment systems to solve micro-efficiency problems by introducing positive incentives in the form of direct financial rewards.

[25] See section 1.6.4.4.

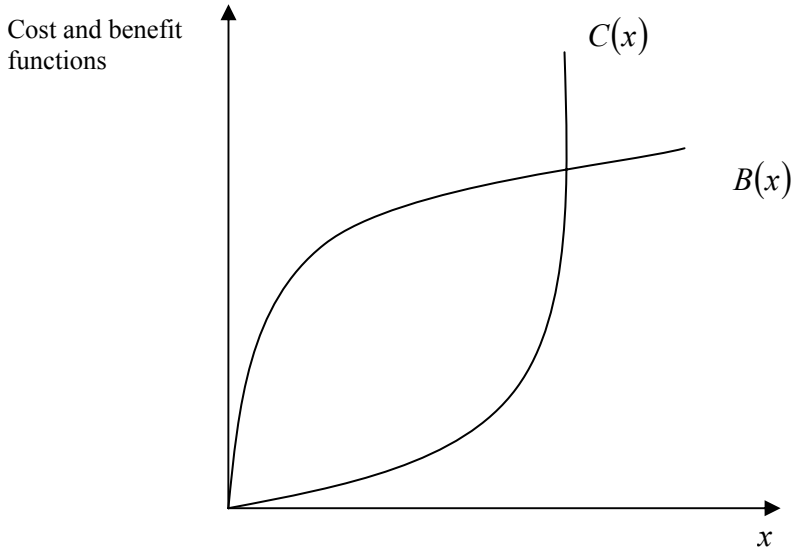


Fig. 14 Physician's cost and benefit functions

The reimbursement rule of these monetary costs is given by the following expression:

$$R + rC(x) \qquad 0 \le r \le 1 \qquad [1.62]$$

i.e. a fixed part R (*lump sum*) that corresponds to the prospective part of the reimbursement plus some amount of cost (r) which is reimbursed by the payer.

A pure retrospective payment system is characterized by

$$R = 0 \qquad \text{and } r = 1 \qquad [1.63]$$

in which all costs are reimbursed whatever the amount of the cost.

In contrast, a pure prospective payment system is characterized by

$$R = EC(x^*) \qquad \text{and } r = 0 \qquad [1.64]$$

In this case, there is no cost reimbursement and the physician receives a positive *lump sum* equals to the expected cost of care (otherwise the physician will made a

loss). Because of the lump sum is paid before the treatment takes place, the physician is left to do whatever he wants.

Presumably, in order to make the provider responsible for 'costs at the margin' r must be somewhat between 0 and 1 and R must be positive. This kind of payment is what Ellis and McGuire (1986) call 'mixed systems' of reimbursement.

Then, the physician utility, which includes monetary e non-monetary terms, becomes

$$U_{ph} = R - (1-r)C(x) + B(x) \qquad [1.65]$$

i.e. the reimbursement the provider receives $R + rC(x)$ less the cost $C(x)$ that he has to pay plus some benefits from providing care. The provider will choice the level of care x in order to maximise his utility. Taking the first derivative with respect to x and setting it to zero,

$$\frac{\partial U}{\partial x} = -(1-r)C(x)' + B'(x) = 0 \qquad [1.66]$$

the physician's optimum is found. The optimal level of care x^* is thus such that

$$(1-r)C'(x) = B'(x) \qquad [1.67]$$

Under a retrospective payment $(R = 0, r = 1)$, the condition above becomes

$$B'(x) = 0 \qquad [1.68]$$

whereas, under a prospective payment $(R = EC(x^*), r = 0)$ is given by

$$C'(x) = B'(x) \qquad [1.69]$$

Fig. 15 shows as, moving from a fully prospective to a fully retrospective payment system, the level of care provided increases along the marginal benefit schedule. In particular, a mixed reimbursement system will provide a level of care between the two extremes x_p and x_r, and for this reason it '*can be more fair to providers than either extreme*' (Ellis and McGuire, 1993). Plotting the level of treatment provided against the reimbursement (i.e. the level of cost-sharing) a supply curve (of health care) is derived (Figure 17).

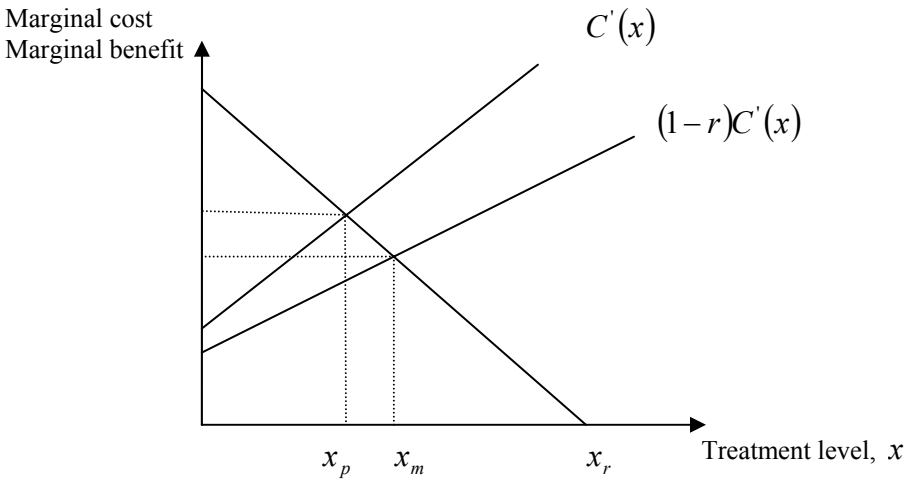


Fig. 15 Treatment level under competing reimbursement systems

The reimbursement rate may also be interpreted as the social price given by the payer for care. The payer can use this supply curve to set the price, i.e. the degree of cost sharing, in order to implement some optimal social level of care (Fig. 16)[26].

[26] Ellis and McGuire (1986) make an additional point that supply side cost sharing has to deal with the moral hazard problem which it may face with in the health insurance. Providing more insurance to the patients means that they are likely to consume much care. The introduction of coinsurance may reduce the patient moral hazard. However, from an insurer point of view this is not good because he is exposing the patient to some risk. In considering health payment systems, it is important to take into account demand and supply side payment incentives, i.e. the insurance contract and payment systems towards the provider and then optimise taking both into account.

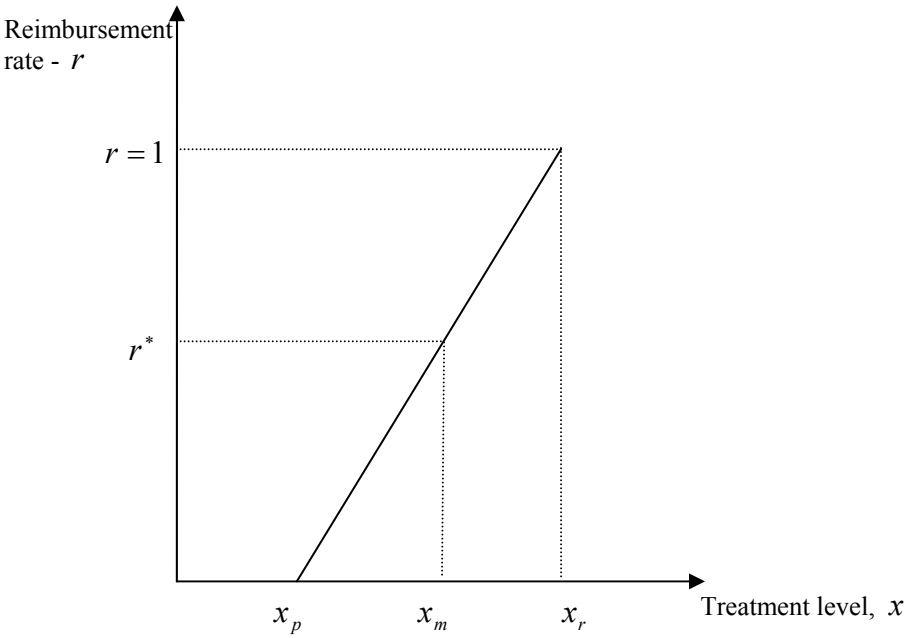
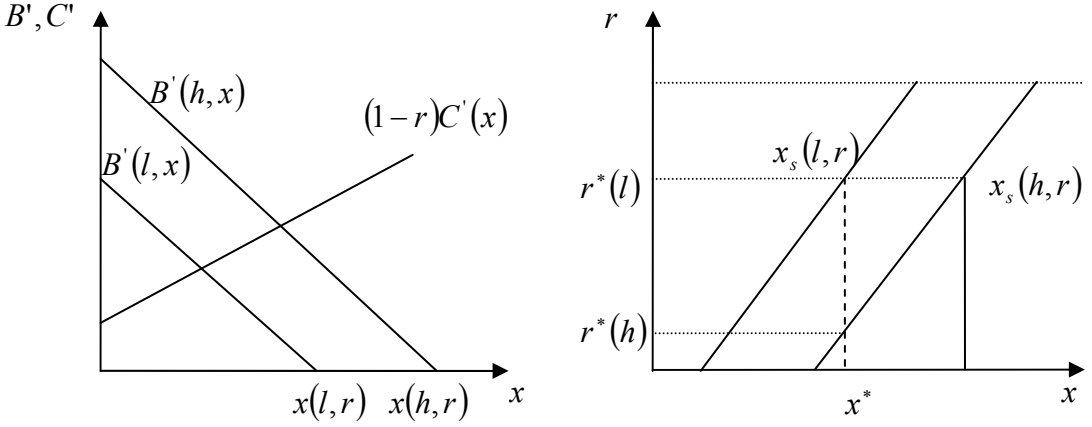


Fig. 16 A supply curve

It is possible to perform some comparative statics with this model. For example, it is possible to see the effect on supply curve from a variation in physician's marginal benefit (Fig. 17). Assume, to this purpose, that somehow there are less altruistic or less status motivated physicians.



Variation in physician's marginal benefit *Effect on suppy curve*

Fig. 17 Some comparative statics (1)

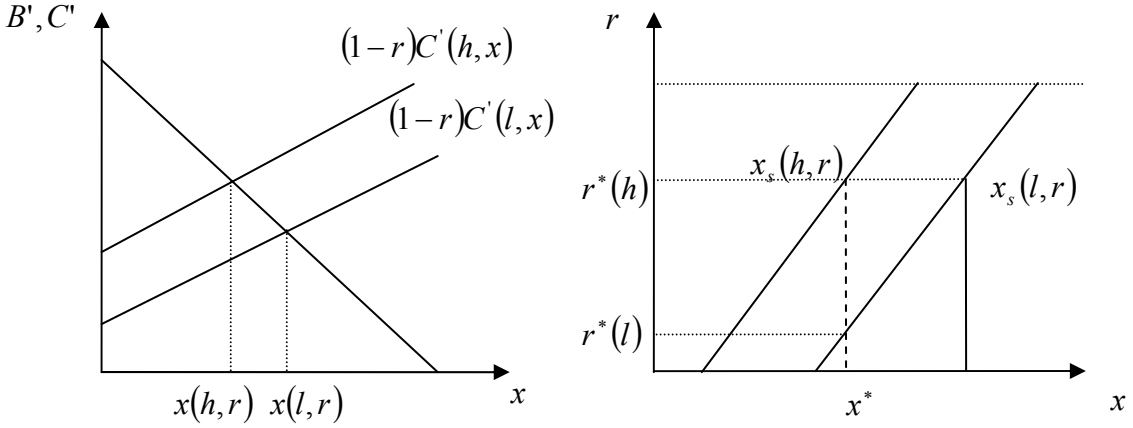
Given the marginal cost of providing care and the level of reimbursement rate, less altruistic physicians will provide a lower level of care

$$x(l,r) < x(h,r) \text{ and } x_s(l,r) < x_s(h,r) \qquad [1.70]$$

If a payer wants to implement a certain level of care, assume x^*, and he does face different physicians in the population (some with a high level of altruism, other with a low level of altruism), then he needs to differentiate the payment system in order to be targeted to the types of physician. This is not a trivial problem from a practical point of view:

$$r^*(l) > r^*(h) \qquad [1.71]$$

A similar analysis can be carried out with respect to differences in marginal costs (Fig. 18). Here, the hypothesis is that physicians can face different types of patients, some less costly others more costly.



Variation in physician's marginal cost *Effect on suppy curve*

Fig. 18 Some comparative statics (2)

Given the marginal benefit of the physician, high cost patients compared with low cost patients will be given a lower level of care. This also translates in a supply schedule where for any level of cost-sharing high cost patients will receive a lower level of care

$$x(h,r) < x(l,r) \text{ and } x_s(h,r) < x_s(l,r) \qquad [1.72]$$

To guarantee the same level of care x^* to both high and low cost patients, we need to target a higher reimbursement rate at those patients which are relatively costly

$$r^*(h) > r^*(l) \qquad [1.73]$$

To summarise the policy implications, if a payer chooses a uniform reimbursement rate across patients and physicians, some variations in practice style may occur when the physicians have different marginal costs (e.g. because of different severity) and/or different marginal benefits (because of different degree of altruism). In this case, the payer needs to provide specific reimbursement rate across physicians to eliminate these variations. This is not

trivial to implement. Moreover at patient level, if a uniform reimbursement is used, a discrimination between patients may take place if they differ in susceptibility to treatment. In this case, the payer has to introduce some patient/case specific reimbursement rate. This is rather easy to arrange.

1.8.3 General issues to be addressed when drawing up doctor's reimbursement

The aim of this section is to discuss further general issues which are relevant in the design of optimal payment system. They include a discussion on risk and the problem of multitasking.

1.8.3.1 Physician participation

This section deals more with the lump sum payment which the physician should receive. From the payer's perspective, there is an incentive to minimise health care expenditure given the level of care to be implemented. In order to minimise health care expenditure, the payer aims to minimise the fixed fee R subject to the physician's participation constraint. Thus, the physician's expected utility must be greater or equal to the reservation utility $\overline{U}$ that is an 'outside' option, i.e. a different occupation or position in the health care system. The physician's expected utility from taking up the job is

$$EU = R - \left(1 - r^*\right)EC(x) + B(x) \qquad [1.74]$$

i.e. the fixed fee the physician receives, plus the benefit from providing care less the expected cost for providing care given the optimal reimbursement rate. Expected costs refer to the fact that there is some uncertainty, as we will see later on.

The payer will reduce the fixed fee paid to the physician up to the point the physician is indifferent between taking up the job and taking the 'outside' option. Therefore, the optimal fixed fee R^* will be

$$R^* = \overline{U} + (1 - r^*)EC(x^*) - B(x^*) \qquad [1.75]$$

This is increasing in reservation utility and physician's share of expected cost and decreasing in the physician's (expected) benefit from treatment.

1.8.3.2 Physician risk

Another general issue that needs to be taken into account is the possible risk that the physician may be exposed to. As a rule, if an agent is risk averse then an increased risk in his income needs to be compensated for by some premium, otherwise the agent will not enter into the contract. Then, the physician needs to be reimbursed for the risk ex-ante in the lump sum part. For convenience, assume that the benefit from providing the optimal of care is zero, i.e. $B(x^*) = 0$.

The effective total reimbursement (i.e. the optimal lump sum plus the retrospective part of the reimbursement) in the presence of risk is given by i.e. the reservation utility plus the risk premium plus the expected cost plus which is born by the physician plus the accounted cost - after the physician has provided the services - which is reimbursed.

$$R^* + r^* \hat{C}(x) = \overline{U} + P + (1 - r^*)EC(x^*) + r^* \hat{C}(x) \qquad [1.76]$$

There could be two types of risk. We refer to financial risk so there are no risks related to potentially adding an adverse outcome with respect to the treatment. First of all, there could be forecast errors. Then the expected cost is not necessarily equals to the real cost given that the optimal level of care is provided but there is some error as captured by α in the formula:

$$EC(x^*) = C(x^*) + \alpha, \qquad \alpha \sim N(0, \sigma_\alpha^2) \qquad [1.77]$$

α is a random variable normally distributed with mean 0 and variance σ_α^2.

Risk implies that in some cases the expected cost is greater than the real cost and in this case the physician is happy because the lump sum he receives is greater

then the cost he bears. Problems arise if the expected cost is below the real cost. In this case the physician faces a loss and, if risk averse, he wants to be reimbursed for this risk. Forecast errors can come from unforeseen complications with treating, unforeseen complications regarding the all patient population, uncertainty regarding medical treatment, etc. Forecast errors introduce risk into the fixed part of reimbursement because the expected cost is part of the fixed component of the reimbursement.

If there is the risk due to forecast errors, then the physician wants to be reimbursed by a higher premium. The payer has an incentives to decrease risk and the premium P by reducing the risky component of the physician's income. As risk arises in the fixed part of the physician's income, the payer wants to reduce the fixed part of the income. Essentially, in this situation, payers will set a higher retrospective part of the reimbursement by introducing a grater degree of cost-sharing ($r > r^*$). This means that the greater part of the physician reimbursement depends on the true cost which is actually realised. There is no longer risk involved for the physician. Physicians have the incentive to provide a greater level of care, since they are reimbursed irrespective of the risk and therefore there will be a over provision of care. Specifically, we can not longer attain the social optimal level of care.

The second form of financial risk takes into account the accounting errors. They introduce risk into the variable part of the reimbursement.

$$C(x) = \hat{C}(x) + \beta, \qquad \beta \sim N(0, \sigma_\beta^2) \tag{1.78}$$

In this case, payers aim to reduce this risk premium by setting a lower rate of cost sharing ($r < r^*$).

In both cases r is distorted away from the optimum r^* and generate a trade-off between cost-sharing and efficiency in the provision of the treatment.

1.8.3.3 Yardstick competition

Another issue is related to some common cost shocks (i.e. epidemics) the physician have to face. In this case, physician's reimbursement needs to be flexible in the face of uncertainty.

Assume that a great number of individual physicians n is faced with common cost shocks (e.g., epidemics). Assume that the ex-post reimbursement to provider i is given by the cost-sharing rate times the cost of the individual physician plus an additional term. The additional term is the residual of the reimbursement rate times the ex-post average cost AC in the industry.

$$r^* C_i(x_i) + (1 - r^*) AC \qquad [1.79]$$

The average cost can not be influenced by the physician and is given by

$$AC = \frac{1}{n} \sum_n C_i(x_i) \qquad [1.80]$$

Ex-post average cost across physicians are constant for an individual physician if n is large. The income of the physician, after the treatment has taken place and the shock has been realised, can be written as

$$\overbrace{\bar{U} + P}^{R} + (1 - r^*)[AC - C_i(x_i)] \qquad [1.81]$$

It is formed by the ex-ante income – the outside utility plus the risk premium – and the ex-post income which depends on the difference between the average cost and the physician's individual cost. In this case, physicians have to bear the actual costs which are not reimbursed. For instance, if a physician bears high costs because of the high level of services he provides, the term in brackets can be negative and, consequently, he is exposed to a loss. Conversely, an efficient physician who takes the cost down would receive some residual.

If all providers choose optimal x^* guaranteed by setting $r = r^*$, the average cost is given with respect to x^*. In this case if there is a common shock the average cost will vary jointly with the physician's individual cost and

$$AC(x^*) - C_i(x^*) = 0 \qquad [1.82]$$

Whenever a common shock there is no longer an income risk. This form of payment mechanism which is labelled yardstick competition eliminates the physician common risk without distorting incentives.

However, this payment system presents some drawbacks. One problem is that yardstick competition does not work under physician specific shocks. For example, assume for some reason a particular doctor faces very expensive patients to treat which is more than the average, then the average cost in the industry will be still very low and the specific physician can not be insured against this risk by yardstick competition. The other potential problem is in a situation in which the number of physicians is relatively small and the industry's average cost will be no longer independent of the cost of individual provider. If the number of providers is small, industry average cost (and reimbursement) may be manipulated by individual or colluding physicians. There will be an incentive to over provide services. Physicians raises their own costs but they will rise also the average cost.

1.8.3.4 The problem of multitasking

Another issue concerns the problem of multitasking. Multitasking concerns with the situation in which the physician has to perform more than one task. Specifically, the problem of *multitasking* refers to the challenge of designing incentives to motivate appropriate effort across multiple tasks when the desired outcomes for some tasks are more difficult to measure than others (Holmstrom and Milgrom, 1991). When *x* is unverifiable, because of asymmetric information between the physician and the payer, a trade-off between effort and service provision arises. If the provider acts as an imperfect agent (i.e. phenomenon such

as dumping, cream skimming or skimping arise) then at reimbursement rate r^* that implements x^* effort is underprovided[27].

Physician offers services x and engage in cost-reducing effort e. In this case, financial cost depends not only on the level of services provided but also on some cost-reducing effort. Cost still increases in the level of services provided and decreases in the cost-reducing effort. The cost-reducing effort is not for free. The physician's utility function is given by

$$\max_{x,e} U = R - (1-r)C(x,e) + B(x) - \psi(e) \qquad [1.83]$$

The first order conditions are

$$\frac{\partial U}{\partial x} = -(1-r)C_x(x,e) + B' = 0 \qquad [1.84]$$

$$\frac{\partial U}{\partial e} = -(1-r)C_e(x,e) - \psi' = 0 \qquad [1.85]$$

The optimal level of effort e^* is given at the point in which the marginal cost from providing effort equals the marginal cost reduction

$$-C_e(x,e) - \psi' = 0 \qquad [1.86]$$

i.e at $r = 0$.

[27] Reimbursement systems affect average resources use at providers and at a system-wide level in three ways: (i) moral hazard (i.e. change the intensity of services provided to a given set of patients), selection effect (i.e. change the tyoe or severity of patients they see), practice-style effect (i.e. change their market share), Ellis & McGiure (1986).

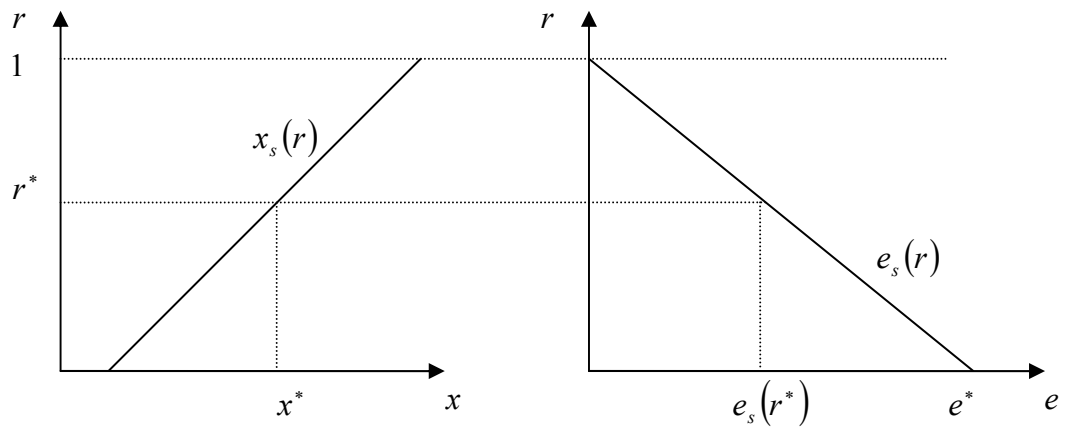


Fig. 19 Supply functions of x and e

At $r = 1$, physicians will provide zero effort. As we gradually reduce the degree of cost-sharing, the physician will have a greater incentive to supply cost-reducing effort. Whenever $r > 0$, there will be an undersupply of effort: a trade-off between efficient effort and service level arise (Fig. 19). This trade-off can be resolved if the level of x can be directly contracted on. Therefore, if and only if the principal can observe x he can replace cost-reimbursement by payment of a price

$$P = r^* C_x\left(x^*, e^*\right) \qquad [1.87]$$

for each unit of x.

The maximisation problem becomes

$$\max_{x,e} U = R + Px - C(x,e) + B(x) - \psi(e) \qquad [1.88]$$

The first order conditions are

$$\frac{\partial U}{\partial x} = P - C_x(x,e) + B' = r^* C_x\left(x^*, e^*\right) - C_x(x,e) + B' = 0 \qquad [1.89]$$

$$\frac{\partial U}{\partial e} = -C_e(x,e) - \psi' = 0 \qquad [1.90]$$

and consequently x^* and r^* are implemented.

But if payments cannot be made contingent on x, e.g. because x is unobservable and then not contractible, the first best is unattainable.

1.8.4 Payment mechanisms and incentives for quality

A payment mechanism is defined as a type of contract among two or more players - patients, physicians and payers – that produces specific incentives for the provision of medical care and minimize the risk of opportunistic behaviour. The aim of this section is to see what incentives for quality arise from different payment systems. We will discuss the influence of payment systems designed by the payer on physician behaviour, highlighting their implications for rational resource use in primary health care. Each of them is likely to have both positive and adverse effects on the nature, quantity and quality of care that a patient receives. We draw the main results from Zweifel and Breyer (1997). In order to build a formal framework, they formulate the following assumptions.

Assumption 1 The physician offers only two different kinds of services whose quantities are M_1 and M_2

Assumption 2 The production function for the *i*th service is given by

$$M_i = f_i(t_i, x_i) \text{ with } \frac{\partial f_i}{\partial t_i} > 0, \frac{\partial f_i}{\partial x_i} > 0 \qquad i = 1,2 \qquad [1.91]$$

where t_i denotes the physician's working time (also effort) and x_i another factor input in the production of the *i*th service. Therefore, only two inputs are considered. The physician buys the input x_i at an exogenous price w in a competitive market.

Assumption 3 The contribution to the patient's health H is a function of both medical services provided,

$$H = H(M_1, M_2) \qquad [1.92]$$

H increases with the provision of both medical services but beyond a certain amount $\left(M_1, M_2\right)$, additional medical services becomes dangerous, causing H to decrease.

Assumption 4 The physician's utility is an increasing function of the revenue Y and the achieved therapeutic success H and a decreasing function of his working hours t.

$$u = u(Y, t, H), \qquad u_Y > 0, u_t < 0, u_H > 0 \tag{1.93}$$

When maximizing his utility, the physician must take three restrictions into account.

Restriction 1 His total working hours are the sum of the working hours used in the production of the two services provided

$$t = t_1 + t_2 \tag{1.94}$$

Restriction 2 Physician's income is given by his non-labor income Y_0 plus the revenue $E(\cdot)$ less costs

$$Y = Y_0 + E(\cdot) - w(x_i) \tag{1.95}$$

Restriction 3 Contribution to patient's health

$$H = H\left[f_1(t_1, x_1), f_2(t_2, x_2)\right] \tag{1.96}$$

The objective function is given by

$$u = \left(Y_0 + E(\cdot) - w(x_i); \sum_i t_i; H\left[f_i(t_i, x_i)\right]\right) \qquad i = 1,2 \tag{1.97}$$

which is maximized by the physician's choice of quantities of inputs t_i and x_i. This maximum depends on the type of the revenue function $E(\cdot)$.

1.8.4.1 Salary system

Salary is the most common system of physician remuneration. Physicians are paid a fixed amount which is decided prospectively for predetermined hours of work. This amount does not depend on the number of patient visits, number of cases, severity of cases, etc. In this case, the revenue formula is given by

$$E = E^* \qquad [1.98]$$

The first-order conditions for a maximum of the physician's utility (interior optimum) are thus ($i = 1,2$)

$$\begin{aligned} \frac{\partial u}{\partial t_i} &= u_t + u_H \frac{\partial H}{\partial M_i}\frac{\partial f_i}{\partial t_i} = 0 \\ \frac{\partial u}{\partial x_i} &= -w u_Y + u_H \frac{\partial H}{\partial M_i}\frac{\partial f_i}{\partial x_i} = 0 \end{aligned} \qquad [1.99]$$

Dividing these equations by each other, we obtain

$$\frac{\partial f_i / \partial t_i}{\partial f_i / \partial x_i} = \frac{-u_t / u_Y}{w} \qquad [1.100]$$

The marginal rate of substitution between the factors of production equals their factor price ratio, which implies that the condition for an efficient production of services is satisfied.

Salary also induces a cost-minimizing combination of services. From the optimality condition [1.99]

$$\frac{(\partial H/\partial M_1)(\partial f_1/\partial x_1)}{(\partial H/\partial M_2)(\partial f_2/\partial x_2)} = 1 \qquad [1.101]$$

The marginal therapeutic success brought about by an increase in the quantity of any one factor x_i is the same for both kinds of services, M_1 and M_2. Therefore, payment by salary induces the physician to achieve the desired therapeutic success in an efficient way.

If $u_H = 0$, the only solution is a corner solution with $M_1 = M_2 = 0$. Therefore, salaried (non altruistic) physicians have the incentives to distort the structure of services by providing excessive referral and prescription in a way that minimize effort. Consequently, under a salary-based remuneration system physicians tend to not deny access and treatment to any patient even thought access may be somewhat inadequate by the limited time that the physician may allocate to his tasks. In addition, since the extent or type of treatment is not likely to alter their compensation, physicians have no incentive to provide excessive treatment. If u_H is still positive but small compared with u_Y and $-u_t$, the optimality conditions [1.99] are only met if the marginal factor productivities as well as the marginal productivities of both services in achieving treatment success are large. As the production function has decreasing marginal returns, this implies that factors are employed but in small quantities and only few services are provided. Thus it is likely that therapeutic success falls short of the level desired by society.
As the physician's income is not linked to the demand of their services, (market) competition has no direct effect.
In the absence of incentives, performance becomes an 'individual' function that is not guided by 'institutional' considerations. Only if the physician is motivated intrinsically or by the status the salary remuneration may still be optimal.

Possibly the biggest advantage in the salary-based remuneration system lies in the ease and simplicity of administration. There are no patient bills to be processed, no patient lists to be prepared, and no case-based groups to be formed. Physicians are treated like other salaried employees, and their service records and payrolls maintained in the same manner. Their promotions are usually time-based, like rest

of the staff. Personnel costs are known in advance, and can be built into any planning exercise of the health department.

To mitigate adverse effects of a salary-based remuneration system the government/payer can offer non-pecuniary incentives to physicians, like awards, favorable posting, status-related designations etc. Even pecuniary benefits can be built into a salary system. One example of such incentives are performance-related financial bonuses. Governments/payers can set in place a system of quality control to monitor and maintain quality levels. Governments/payers can improve monitoring to ensure greater availability of physician time.

1.8.4.2 Capitation system

Compared with fixed salary, the capitation system does not depend on input variables but on variables positively related to output. For example, salary can be linked to the number of registered patients (P), which would presumably reflect the degree of success of the physician's treatment [$P(H)$ with $P'(H) = dP/dH > 0$]. In this case, physicians are paid a fixed fee (q) for each ensured enrollee in their list to cover a specified level of health care and offer a defined package of services, for a specified period of time. Thus the revenue function is given by

$$E = qP(H) \qquad [1.102]$$

$P(H)$ can be regarded as the patient demand.

By using physician time as the only input, we simplify the production function as following:

$$M = M(t) \qquad [1.103]$$

where M is the total amount of services to be distributed among P patients.

It is assumed that the average degree of success in treatment depends on the quantity of services per registered patient M/P:

$$H = H\left(\frac{M(t)}{P(H)}\right) \qquad [1.104]$$

Hence, the physician's objective function is

$$u = u\left\{Y_0 + qP(H), t, H\left[\frac{M(t)}{P(H)}\right]\right\} \qquad [1.105]$$

By differentiating $H(\cdot)$, we obtain the effect of an extension of physician working hours on the average degree of success in treatment

$$H'(t) = \frac{\partial H}{\partial(M/P)} \frac{P(H)M'(t) - M(t)P'(H)H'(t)}{P^2} \qquad [1.106]$$

Solving this equation for $H'(t)$ and denoting $\partial(H)/\partial(M/P)$ by $H'(M/P)$, we obtain

$$H'(t) = \frac{P(H)H'(M/P)M'(t)}{P^2 + H'(M/P)M(t)P'(H)} > 0 \qquad [1.107]$$

Hence, we can write the first order condition for utility-maximizing working hours as

$$\frac{du}{dt} = \left[u_Y qP'(H) + u_H\right]H'(t) + u_t = 0 \qquad [1.108]$$

Even thought $u_H = 0$, i.e. the physician is not interested in therapeutic success, a positive H is reached. At the optimum, the disutility due to leisure time sacrificed must be equals to the marginal utility of additional income due to additional

registrations, which in turn are the result of an increased therapeutic success due to an increased time (effort) devoted.
By choosing an appropriate level of q, it is possible to achieve any level of treatment success desired by society H^*. The optimal level of q is given by solving the first order condition for q and by inserting H^* (and the corresponding quantities t^* and Y^*).

Physicians under capitation are usually responsible for all costs of providing the full package of treatment, including diagnostic tests, specialist consultations, and some minor, ambulatory surgery. Physicians thus receive a fixed amount per enrollee, and after meeting all costs of treatment, retain the surplus as their income. Since the compensation package is decided prospectively, physicians can maximize the difference between their earnings and costs by simply keeping costs down.

The incentives properties of capitation have been widely studied (Zweifel and Breyer 1997 section 7.3; Gravelle 1999; Chalkley and Malcomson 2000; McGuire 2000). Since they face both the monetary and effort cost of providing care, they have a strong incentive to reach an efficient mix of resources. In order to keep costs down physicians invest in preventive medicine, tend to enroll relatively healthy patients, who are likely to require less frequent and less costly treatment, to limit the quantity of services provided to the patient, as this would reduce their operating costs, to refer patients to next higher levels of care, such as to specialists and hospitals, so as to save own operating costs. The patient is likely to receive only those services and interventions that are necessary. Over-treatment and unnecessary interventions are unlikely in this system. Patients requiring many and complex treatments may be excluded from the enrollment lists of physicians receiving capitation-based remuneration. Patients may potentially receive less than optimal care, since the physician has an incentive to keep costs down. Both quality and quantity of care may thus be compromised.
In many capitation-based settings, it is not easy and straightforward for the patient to change her physician.

The capitation system requires little administration, as there are no patient bills to be processed. Once the lists are finalized, computing an individual physician's reimbursement does not require many calculations. Health expenditures under this system are entirely predictable, and there is good control over costs. However, governments and payers may incur high administrative costs in preparing and maintaining the list of enrollees. High administrative costs are also associated with negotiating contracts, setting capitation rates and formulas, and monitoring physicians to ensure that appropriate levels and quality of care are being provided to the patients.

To mitigated the adverse incentives facing the physicians governments/payers can mandate open enrollment, such as across a given geographical or administrative area. This prevents the physicians from selecting the relatively healthy patients into their pool. At the same time, the potential risk to the physicians is also minimized, since spreading the base from which to select enrollees reduces the chances of relatively unhealthy members dominating the pool. Governments/payers can encourage competition among physicians. If the patients have the choice to periodically select their physicians, the potentially adverse consequences for quantity and quality of care can be overcome.

Governments/payers can define the package of services to include many (not necessarily all) such services that are usually provided at a next higher level. Thus, if family care physicians are required to provide diagnostic services as well, there will be fewer incentives for the physician to refer patients for additional tests. Similarly, if visits to a specialist are covered by the physician as part of the capitated services provided, there will be fewer incentives to refer patients to specialists

Governments/payers should set in place a system of quality control and monitoring so as to mitigate the adverse incentive that the physician faces of compromising on quality.

The capitation formula is usually based on variables on which reliable information is available and easy to get. Variables commonly used in capitation formula are: number of individuals, age, gender, marital status, socioeconomic conditions,

urbanization, mortality and morbidity, number of individuals disabled, number of individuals under rehabilitation, number of individuals with substance abuse problems, epidemiological indices.

Countries that follow a capitation-based system for remuneration of health care personnel are: Hungary, Ireland, Italy, Holland, Croatia (private sector). In some countries the principle of capitation is used to define provincial and district health care budgets. Fund allocation across different administrative areas follows use of capitation formulas that typically include demographic and epidemiological indices.

1.8.4.3 Fee-for-services system

Under this system the physicians are paid according to the number and type of different services provided to the patient. A specific price p_i ($i = 1,2$) is set for each service i, which may be fixed or variable. In the fixed system, the physician and the payer agree to a schedule of rates at the beginning of the year, and all bills within the agreed time-period are based on this schedule. In the former case, the revenue function is given by

$$E = \sum_i p_i M_i = \sum_i p_i f_i(t_i, x_i) \qquad i = 1,2 \qquad [1.109]$$

The first-order condition for an (interior) utility maximum

$$\begin{aligned} \frac{\partial u}{\partial t_i} &= \left(u_Y p_i + u_H \frac{\partial H}{\partial M_i} \right) \frac{\partial f_i}{\partial t_i} + u_t = 0 \\ \frac{\partial u}{\partial x_i} &= \left(u_Y p_i + u_H \frac{\partial H}{\partial M_i} \right) \frac{\partial f_i}{\partial x_i} - u_Y w = 0 \end{aligned} \qquad [1.110]$$

Dividing these equation by each other, we obtain

$$\frac{\partial f_i / \partial t_i}{\partial f_i / \partial x_i} = \frac{-u_t / u_Y}{w} \qquad [1.111]$$

The marginal rate of substitution between the factors of production equals their factor price ratio, which implies that the condition for an efficient production of services is satisfied.

From the first order conditions we obtain

$$\frac{(\partial H / \partial M_1)(\partial f_i / \partial x_i)}{(\partial H / \partial M_2)(\partial f_2 / \partial x_2)} = \frac{(p_2 u_Y + u_H \partial H / \partial M_2)(\partial H / \partial M_1)}{(p_1 u_Y + u_H \partial H / \partial M_1)(\partial H / \partial M_2)} \qquad [1.112]$$

The right-hand side of equation above must equal 1 to achieve a given therapeutic success using the minimum possible amount of services. This is satisfied only if fees reflects exactly their marginal productivities in terms of therapeutic success,

$$\frac{p_1}{p_2} = \frac{\partial H / \partial M_1}{\partial H / \partial M_2} \qquad [1.113]$$

In general, this condition is not satisfied. Then, physician's services are not produced at minimum cost. Physicians have strong incentives to over-provide and over-use those services on which they receive a positive mark-up, i.e. the fee exceeds the marginal cost (Zweifel and Brayer 1997 section 7.3; McGuire 2000). Patients requiring many and complex treatments are not likely to be excluded from receiving care. This is likely to inflate total spending on health care. Physicians have a strong incentive to induce demand. This phenomenon, also known as "supplier induced" demand, is observed to be widespread in countries where physician remuneration is based on fee-for-service (see section 1.4.3). Over-providing may imply a direct reduction in quality and patient welfare. For instance, a purely income-maximizing physician could prescribe a service up to a point at which the marginal effect on health outcomes becomes negative. Even if the benefit on health is positive, it may become small enough to be outweighed by the patient's time or travel cost or some other disutility associated with treatment. Besides, unless the fee reflects the patient's 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 Brayer 1997, section 7.3). Finally, a physician may reduce his effort by investing less time in the provision of each services leading to a deterioration of quality along with a higher volume in the quantity.

Under the fee for service system, reimbursement is made on the basis of services actually provided. Physicians keep a detailed record of quantity and types of services provided, and send this information periodically to the government. Each item of the bill is scrutinized at this level before being cleared for payment. Thus, the administrative costs of such a system are likely to be high for both, the physicians as well as the government.

Fee for service system is likely to increase the production and the productivity in the health care system, but usually at a cost. Experience in countries where this system is popular has shown that rapid increases in costs is a common feature.

To mitigated the adverse incentives facing the physicians, governments/payers can increase control over utilization of health care. However, this involves high monitoring costs. Governments/payers can introduce cost sharing through a system of co-payments. Governments/payers can introduce upper limits of reimbursement, thereby fixing a maximum level of services that would be supported.

Countries that follow a fee-for-service system for remuneration of health care personnel are: Germany (private physicians and dentists), Belgium, Czech Republic, France, Switzerland, Ireland (private sector), Holland (private sector), Bulgaria (private sector), Greece (private sector), Slovenia (private sector), Turkey (private sector).

It is relatively easy and straightforward for the patient to choose and change physicians at any point.

1.8.4.4 Mixed systems

In an effort to optimize efficiency, equity and quality of care, many countries have experimented with mixed systems of remuneration. These systems use more than one method of payment in an attempt to draw on the best that each method has to offer, and at the same time, mitigate the worst of each.

For example, a mixed system employing both capitation and fee-for-service remuneration methods provides benefits of both the methods and has the potential of mitigating the adverse incentive the physicians have in a pure capitation method of producing and delivering less than optimal quantity of care. Combining salary and fee-for-service introduces financial incentives for physicians that are not available in a pure salary system.

Previous economic arguments for mixed payment rely on fighting risk selection and quality stinting while maintaining some incentive to control cost (Ellis and McGuire, 1990; Ma, 1994; Newhouse, 1996; Ma and McGuire, 1997; Pauly, 2000; Eggleston, 2000; Newhouse, 2002). We present some results from Ma's model (1994). The model applies in the context of hospitals but it is equally valid in the context of primary care.

First, it is assumed that the consumers' choice of the provider depends on the quality of care (i.e. full insurance). Second, the provider maximizes its net profit and is a perfect agent (i.e. no dumping, no cream skimping). Third, payer values provider profit as highly as patient welfare then he is concerned uniquely with efficiency issues. The model functions are described below.

The provider's net profit function under retrospective payment is given by

$$\pi = \left[\left(c\left(t_{1,}t_2\right)+m\right)\mu\left(t_1\right)\right]-c\left(t_{1,}t_2\right)\mu\left(t_1\right)-\gamma\left(t_1+t_2\right) \qquad [1.114]$$

We can re-write it as

$$m\mu\left(t_1\right)-\gamma\left(t_1+t_2\right) \qquad [1.115]$$

The provider's net profit function under prospective payment is given by

$$\pi = p\mu(t_1) - \underbrace{c(t_1, t_2)\mu(t_1)}_{Cost-\text{int}ernalisation} - \gamma(t_1 + t_2) \qquad [1.116]$$

The regulator's objective function is

$$B_r(t_1, t_2) = \underbrace{W(t_1)}_{ConsumerBenefit} - \underbrace{[c(t_1, t_2)\mu(t_1) + \gamma(t_1 + t_2)]}_{TotalCostOfThe\,\text{Pr}oduction} \qquad [1.117]$$

where

t_1 is the effort a provider can direct to quality enhancement;

t_2 is the effort a provider can direct to cost reduction;

$c(t_1, t_2)$ the unit cost of treating a patient $c_1 \geq 0$ and $c_2 < 0$;

$\mu(t_1)$ provider's demand when he provide services at quality t_1;

$c(t_1, t_2)\ \mu(t_1)$ total cost (convex function);

m nonnegative margin to motivate the provider to supply effort and quality enhancement;

$\gamma(t_1 + t_2)$ total disutility of the provider imposed by the efforts t_1 and t_2;

$W(t_1)$ gross benefit produced by the provider.

The first order condition from the provider maximisation problem under retrospective payment is

$$m\gamma'(t_1) = \mu'(t_1) \text{ given } m \qquad [1.118]$$

If t_1^+ is the maximiser of the payer's objective when t_2 is set at its minimum, then it can be implemented by setting

$$m = \gamma'(t_1^+) / \mu'(t_1^+) \text{ where } t_1^+ > t_1^* \qquad [1.119]$$

The first order condition from the provider maximisation problem under prospective payment is

$$pW'(t_1) = \mu'(t_1) \text{ given } p \qquad [1.120]$$

If t_1^* and t_2^* are the maximisers of the payer's objective, then they can be implemented by setting

$$p = W'(t_1^*) / \mu'(t_1^*) \qquad [1.121]$$

Hence, the optimal cost reimbursement policy can induce a constrained efficient quality effort (in the sense that it is set at the minimum). The prospective payment system can achieve the efficient allocation of effort between cost and quality. We can draw from these solutions a first result. The prospective payment system is the best payment mechanism: it unambiguously represents an improvement over retrospective payment. '*However, it may not in fact be efficient to treat all those patients who wants treatment at the quality offered when patients do not themselves pay for treatment. Moreover, if the provider's services are used to capacity, the number of treatments may necessarily be less than the demand for them and the provider may not incur the cost of increasing quality in order to increase demand*' (Chalkley and Malcomson, 1998).

In presence of dumping a reimbursement rule must be introduced to implement efficient effort. Under a reimbursement rule a provider's payment will be increased to cover costs whenever these costs are above a preset level c^*; otherwise the provider is fully responsible for his costs (Fig. 20).

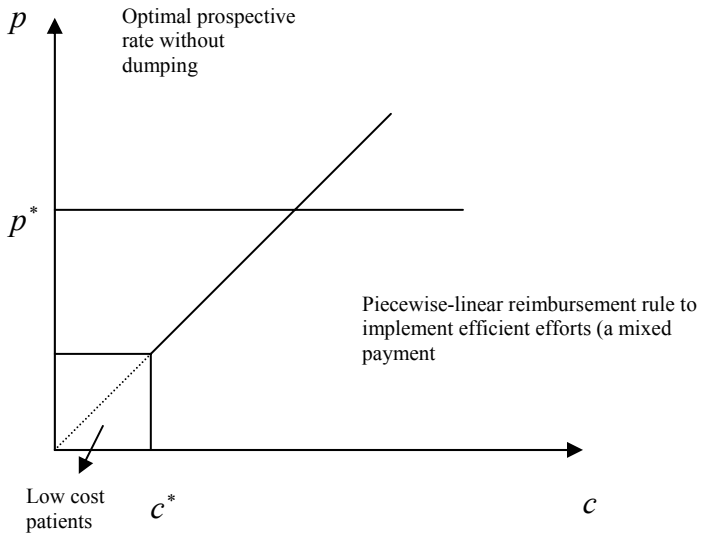


Fig. 20 The reimbursement rule (Source, Ma 1994)

Relaxing the assumption that the physician is a perfect agent, we can make the following statements (second result). The optimal cost reimbursement policy can *again* induce a constrained efficient quality effort. Cost reimbursement can avoid dumping. As for prospective payment system, there is no a priori reason that it must perform better than retrospective payment. The implementation of the efficient allocation of efforts, t_1^* and t_2^* can be achieved by a piecewise linear-reimbursement rule that can be interpreted as a mixture of fully cost-based and fully prospective payment systems.

To sum up, under the assumption of perfect agency, Ma (1994) model shows that in terms of efficiency the prospective payment system is the best payment system. However, relaxing that assumption mixed payments provide other benefits related to the distortion due to the agency problem. They avoid dumping (relative to a pure prospective system) as we have shown; reduce provider's risk in the net revenue in the case of heterogeneity (relative to a pure prospective system) and, finally, discourage low value admissions that are profitable for provider (relative to a pure prospective system).

Mixed payments have also been recognized for having the advantage of balancing incentives for quality effort across contractible and noncontractible dimensions of quality in a context of multitasking (Eggleston, 2005). We will discuss these new results in the next sections.

1.8.5 Hard regulation - Quality-Related Performance Pay

Under existing payment systems, physicians usually receive the same payment regardless of the quality of care provided to patients. Quality incentives are generally implicit in the payment system as we discussed above. To alter this equation, quality-related performance pay has been introduced in a number of health plans[28]. The regulator establishes a link between quality and the physician's income. Quality-based payment also appears feasible within a variety of underlying payment systems. Each of the principal types of payment system - budget transfer (or salary), capitation, and fee-for-service - can be adapted to incorporate financial incentives for quality. Typically, pay for performance programs refer to payment arrangements that offer financial rewards to physicians meeting specific goals, such as provision of certain preventive care, patient satisfaction, acquisition of information technology, and cost containment. In general, the payment must be linked to an observable and verifiable signal correlated with the physician's effort. This signal may be relate to a process or outcome measure (Zweifel and Breyer, 1997 section 8.3). However, most of these programs are in the early stages of trial, evaluation, and adjustment. Although there is intense interest in about pay-for-performance programs among several policy makers and payers, the research published on pay-for-performance in health care is limited. One controversy is whether to compensate physicians according to attainment of a predetermined level of performance or according to improvement. The first way is common to the majority of pay-for-performance

[28] For example, the new general practitioner contract in the UK includes 146 performance measures across seven areas of practice, affecting about 18% of practice earnings. The few US initiatives in this area do currently use only modest financial incentives to reward measured dimensions of physician and hospital performance (Rosenthal et al., 2004; Strunk and Hurley, 2004).

programs. However, physicians that have historically performed above the targeted level are likely to have no incentives to improve because they can receive the bonus simply for maintaining their current level of performance. Moreover, physicians whose performance is initially much below the target may be less stimulated to improve their performance when the target seems difficult to reach. On the other hand, paying for improvement may fail to reward the best physicians for whom improvement appears difficult to attain because of ceiling effects. Another possibility is given by the use of past performance as a benchmark. In this circumstance, the ratchet effect may be a concern. Whenever the payer observe a good performance in one period and 'ratchet up' the benchmark for the next period, then good performance in the first period is punished. The prospective GP expects this, and he will have the incentive to under-perform to benefit from a lower future performance target. As a consequence of this gaming, under-performings becomes andemic to the system (Kuhn, 2003). A problem which can arise under a pay for performance system is also the 'hold up' problem. This consists of the incentive of the payer to announce a satisfactory level or remuneration before the future GP invests in medical capital, and to break this promise later to extract the rent, once the GP has made their investments in human and technological capital. On the other hand, GPs rationally anticipate being 'held up' and they will under-invest or not enter the profession. Hence, a sub-optimal level of quality is provided. To avoid this problem, 'the payer should acquire a credible reputation for maintaining fair remuneration' (Kuhn, 2003 - p. 98).

The major problem recognised by the general literature on incentive is that payment for performance leads to difficulties in presence of multitasking. Accurate measure for provider actions that promote quality are extremely difficult to quantify. The signals are only imperfectly correlated with the true effort and are also subject to random influences and manipulation. When process and outcome measures are subject to uncertainty, a bad outcome may still arise from a good level of effort and *vice versa* (Zweifel and Breyer, 1997 section 8.2). In addition, quality is multidimensional (see section 1.2). As Newhouse (2002) notes, "payment on specific process measures of quality, such as beta-blockers after a heart attack, can distort resource allocation to the measured areas and away from

unmeasured areas (the multitasking problem or 'teaching to the test'). It is therefore hard to know whether on balance patients are better off" (p. 203). Smith and York (2004) identify some potential risks concerned with the new UK contract for general practitioners, including potential stinting on services that are not rewarded (e.g. aspects such as continuity and advocacy) or are only partially rewarded because appropriate measures seemed lacking (e.g., mental health). As a result, because of multitasking and since quality is rewarded only partially or metrics are imperfect payers should use pay-for-performance carefully. In general, the less precise the measure of performance, the lower-powered pay-for-performance incentives should be. There is also a concern about whether incentives should be team-based and related to the outcome or whether they should be target at the individual level (Ratto et al, 2001; see the following sections: 1.6.5.1 and 1.6.5.2).

1.8.5.1 The problem of multitasking. Further considerations

When quality signals used as a basis for pay-for-performance are imperfect – with some important aspects of quality unmeasured and therefore unrewarded – pay-for-performance incentives *and supply side cost sharing* should *both* be used. The problem of multitasking further reinforces the argument in favour of mixed payment systems. Specifically, mixed payment reduces distortions in effort allocation when pay-for-performance is imperfect by stimulating noncontractible dimensions of quality that complement treatment (Eggleston, 2005). The following section uses a simple model to show that mixed payment helps pay-for-performance avoid providers for skimping on noncontractible dimensions of quality. It is drawn from Eggleston (2005). Physicians are assumed to be 'benevolent' i.e. they take into account the patient benefits from treatment (Chalkley and Malcomson, 1998) which depends on resource use m and physician quality effort, e and it is increasing in both of them. The payer's objective is to maximize quality less the cost of paying the physician. The cost of care m is an imperfect measure of performance

$$m = \mu e \quad [1.122]$$

where

$$\mu \sim N\left(1, \sigma^2\right) \quad [1.123]$$

may measure variation in case-mix not captured by the risk adjustment system. The provider bears a disutility for providing effort $c(e)$ increasing and convex in e. Provider net revenue per patient, π, consists of three components: pre-payment R; pay-for performance incentives ρ; and supply-side cost-sharing for each service defined by the fraction of costs that the payer reimburses r ($0 \le r \le 1$):

$$\pi = R + \rho e + rm - m = R + pe + (1 - r)m \quad [1.124]$$

and the utility maximisation problem becomes

$$\underset{m,e}{Max}\left[(av(m,e)) + R + pe + (1 - r)m - c(e)\right] \quad [1.125]$$

where $\alpha v(m,e)$ is the degree of altruism; the higher is α the more benevolent the physician is. The utility is assumed to be strictly concave. The first-order conditions define spending m^* and effort e^* as functions of provider benevolence and the payment parameters

$$\alpha v_m\left(m^*, e^*\right) = s \quad [1.126]$$

$$\alpha v_e\left(m^*, e^*\right) + \rho = c'\left(e^*\right) \quad [1.127]$$

where $s = 1 - r$ is the reimbursement rate. Their values deviate from first-best. The first-best is given when the regulator can directly contract for quality and spending, i.e.

$$\underset{m,e}{Max}\left[v(m,e) - m - c(e)\right] \quad [1.128]$$

implying

$$\alpha v_m\left(m^*, e^*\right) = 1 \qquad [1.129]$$

$$\alpha v_e\left(m^*, e^*\right) = c'\left(e^*\right) \qquad [1.130]$$

When the physician do not care totally patient benefits ($\alpha < 1$), he has an incentive to provide less effort and spending more ($\alpha v_e < v_e; \alpha v_m < v_m$) unless compensated for the disutility and the resource use. Low supply cost sharing s and positive pay-for-performance ρ lower the physician's marginal cost of spending and arises the physician's marginal benefit from quality effort. This is simple to show by making equivalent the first order conditions when the payer can directly contract for spending and quality and when he cannot.

$$(1-\alpha)v_m = 1 - s = r \qquad \text{if } \alpha = 1 \text{ then } r = 0 \text{ ; if } \alpha < 1 \text{ then } r > 0 \qquad [1.131]$$

$$(1-\alpha)v_e = \rho \qquad \text{if } \alpha = 1 \text{ then } \rho = 0 \text{ ;if } \alpha < 1 \text{ then } \rho > 0 \qquad [1.132]$$

Then 'supply-side cost sharing can be set appropriately to promote cost control, while direct rewards for quality prevent cost control from adversely impacting quality' (Eggleston, p. 217). Provided that higher-quality care implies higher costs and performance measures are imperfect, pay-for-performance and mixed payment should be viewed as complements, not substitutes.

1.8.6 Team-based compensation

Working in teams enables GPs to take advantage of economies of scale (Pauly, 1996), smooth work schedule, and internalise referrals (Gaynor and Gertler, 1995); to insure themselves against idiosyncratic shocks to human capital (Gaynor and Gertler, 1995; Lang and Gordon, 1995). For doctors, idiosyncratic shocks include difficulties in fee collection, reputational damage due to malpractice, variations in insurance coverage, claims denial, shifts in demands for specialties, as well as luck.

Under a team performance plan, each GP receives an equal share of the net earnings of the group, regardless of his individual output and effort. In this context as compensation moves away from individual performance the problem of free riding becomes a concern (e.g., Alchain and Demsetz, 1972; Newhouse, 1973; Holmstrom, 1982)

Hence, it is essential to introduce some mechanisms to mitigate the trade-off between the advantages and the loss in output from diminished incentives due to the free riding on the effort. Traditionally, the literature on team-based incentives predicts that explicit contractual sanctions such as monetary penalties or the treat of expulsion are needed in order to mitigate the free riding on the effort in a team. But another strand of literature argues that also implicit sanctions arising from non contractual and informal interactions that occur between members of work groups are able to discourage such behaviour. The sociology of medical groups uses the term 'group sociology' to refer to these type of interaction. These interactions can take the form of activities (mutual help, mutual monitoring) and psychological experiences (guilt, envy, shame, greed, peer pressure). In particular, three types of informal interactions among physicians have been identified: 1. the intra-group income comparisons that lead to income norms; 2. the intra group effort comparisons and mutual monitoring that result in effort norms; and 3. mutual help activities. According to Kandel and Lazear (1992) and Encinosa et al. (1997) small groups produce sufficient peer pressure to offset free-riding. Che and Yoo (2001) show that when productive interdependence is high, team -based compensation schemes increase members' incentives to monitor each other and increase the power of peer sanctions. The effectiveness of mutual monitoring is expected to increase in the homogeneity of team members (Kandel and Lazear, 1992; Encinosa, et al, 1997) and decrease in very large firms because shirking of the monitoring task will dominate the effects of peer pressure.

1.9 Values and norms as a source of quality incentive

Our discussion has shown that payment systems do impact on the behaviour of physicians. But physicians undoubtedly have motivations other than personal financial gain. Their professional behaviour is also driven by important social and professional norms, altruism and intrinsic motivations. Indeed, an important strand of the literature in economics, sociology, and psychology addresses these aspects and their relationship with economic incentives. For example, McGuire (2000) have addressed the net patient health benefit as an important element driving the physician decision-making process on quantity of output and quality (physician effort) as we discussed above. However, the impact of monetary incentives on intrinsic motivation has not been modelled theoretically in health care. Also the definitive direction of the impact is an empirical question, which to our knowledge has not been investigated in the health services research literature. A priori, a strong system of ethics may attenuate, or totally remove, the incentives to provide an opportunistic behaviour arising from some payment systems. Conversely, external incentives may positively ('crowd in') or negatively ('crowd out') impact on doctor's moral motivation.

According to Frey (1997) the use of external rewards might actually enhance intrinsic motivation if the economic incentive is viewed as legitimating or enhancing internal or professional norms. On the other hand, if the external financial incentives are a substitute for intrinsic motivation, the use of financial incentives might "crowd out" or diminish the strength of internal motivators. The psychological processes underlying the phenomenon have been explained in two ways. Firstly, external incentives may impair self-determination, resulting in a loss of professional autonomy. Secondly, they may damage self-esteem, resulting in the perception that professionalism is no longer valued. Crowding out seems to be stronger when external incentives are linked to perceived regulatory activity and managed in a bureaucratic context by people unknown to the recipients of the incentives. More mechanical tasks seems to be less likely to be crowded out than creative ones. This might explain the support for incentives from those who are inclined to focus on the technical aspects of delivering care, and the antipathy of others who focus on the "art" of clinical practice.

From a social psychological perspective there is evidence, including a meta-analysis of 128 experimental studies (Duci, 1999) that "crowding out" is a real phenomenon. The literature of organisational sociology has recognised for more than half a century that incentivisation of rule governed behaviour is likely to lead to ‘goal displacement’ in which rule following becomes a means to an end other than that intended by the designers of the system. This observation has been drawn upon in more recent sociological writing about the displacement of trust and moral motivation brought about by the current emphasis on "managing" the performance of health and social care professionals (Marshall and Harrison, 2005).

It is essential that we develop a deeper understanding of the relationship between incentivised and non-incentivised professional work. The new UK GP contract provides a case in point. On the one hand, the incentivisation (agreed with the profession itself) of indicators about the routine treatment of single chronic conditions does make sense in the context of the evidence. On the other hand, it seems possible that, as an increasing proportion of total GP work is incentivised, the risks of crowding out of motivation to perform the non-incentivised more complex or simply caring tasks is increased. This is an area of policy that really does need to be underpinned by high quality evidence.

Appendix 1

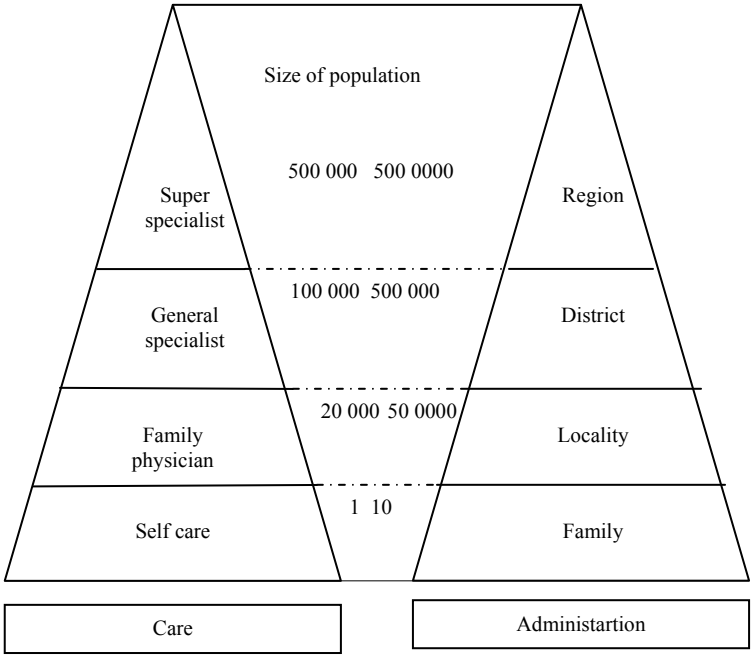


Fig. 21 Levels of care and administration (Source: Fry and Sandler, 1993)

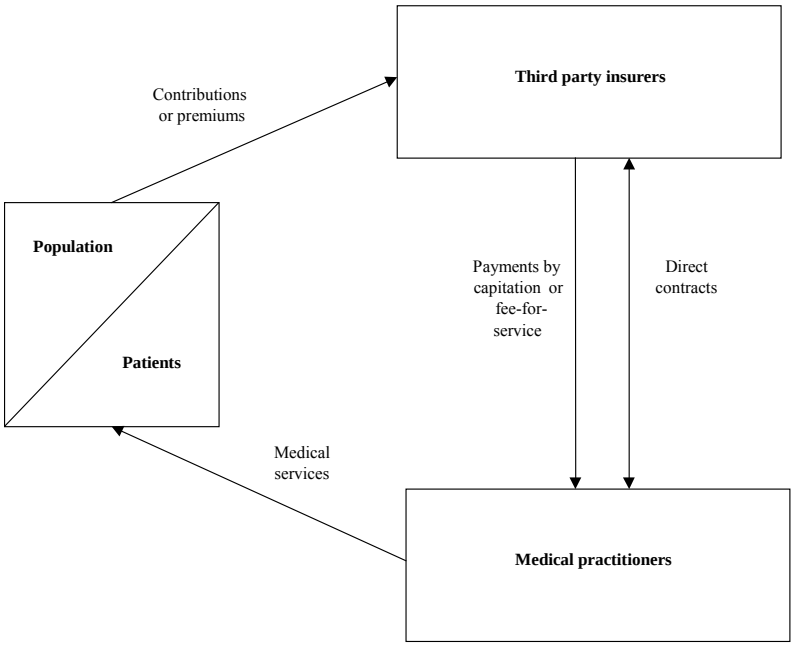


Fig. 22 Contract approach
Source: Bickerdyke et al. (2002) - adapted from OECD (1992)

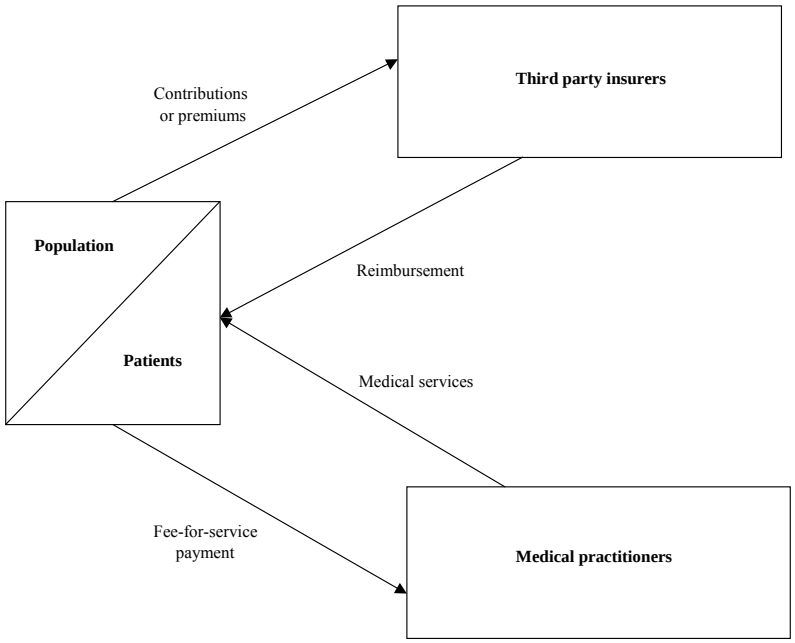


Fig. 23 Reimbursement approach
Source: Bickerdyke et al. (2002) - adapted from OECD (1992)

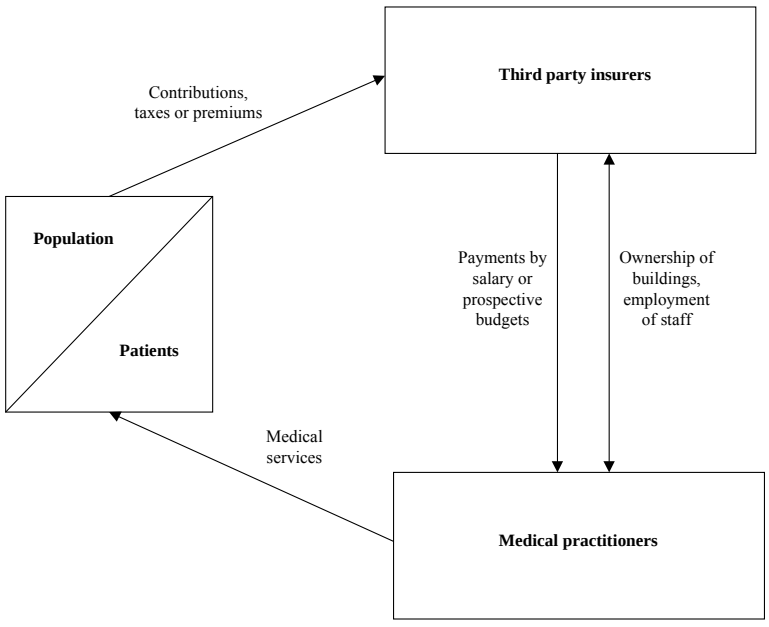


Fig. 24 Integrated approach
Source: Bickerdyke et al. (2002) - adapted from OECD (1992)

Tab. 4 Primary Health Care: Basic Characteristics. (Source: CESifo Dice, 2000 updated through WHO, 2006)

	Organization of PHC	**Choice**	**Gatekeeping and referring**	**Payment at the point of delivery**	**Self-participation**	**Payment method**
Austria	PHC is provided by private GPs and specialists Indirect provision (contracted)	Patients can choose their GPs from among those who have contracts with their health insurance fund.	GPs refer patients to specialists and hospitals. No gatekeeping	Ambulatory care is free at the point of delivery for all patients with general insurance.	Farmers and self-employed people have to pay 20% of the physician's bill. 20% of the population pays 10% or 20%	Fee-for-service. Doctors are paid a standard fee per quarter irrespective of the number of consultations per patient, and additional fees for particular services.
Belgium	Tasks of primary and secondary care are not well defined - no referral is needed to gain access to specialists services or hospital, so many specialists provide primary care. Indirect provision (contracted)	Patients' total freedom of choice (considerable competition between physicians).	Blurred boundaries between GP and specialist - GP has no gatekeeping role.	Advanced payment due insured against compensation for expenses.	10% - 30% (The self-employed pay full cost).	Physicians are paid on a fee-for-service basis.
Denmark	PHC (for residents older then 16) is based on choosing between two health plans :for Group-1 patients access to a GP is free at the point of use, the GP then acting as a gatekeeper. Group-2 patients can visit any GP or specialist without referral but have to pay part of the treatment/consultation costs. In 2002, only 1.7% of the population opted for Group 2, owing partly to the extra costs involved and partly to general satisfaction with the GP referral system. Indirect provision (contracted)	Free choice and patients may change their GP once in six months (Group 1 patients: 98% of population). Absolute free choice (Group 2 patients: 2% of population).	GP has a strong gatekeeping role to the rest of the system. No referral needed for a specialist care.	Free physician and hospital care. Co-payments for all medical services except medical care.	No	28% capitation (flat fee) 63% fee-for-service 9% allowances

Finland	Municipal health centres are providing a full range of primary care service staffed with physicians, nurses, dentists. Direct provision (employed)	Limited free choice (in public healthcare service). Free choice (in insurance companies).	GPs have an important role, since they are involved in all the services rendered by the centres.	Communes decide whether or not to charge for services and to set the level of charges (up to a maximum set by government).	€ 8 for three consultations or € 16 annual fee. By insurance companies: 40% of basic payment. Cost sharing of € 0.2.	Physicians working in the health centres are paid by a combination of basic salary, capitation fee, fees for services provided and 5% local allowance.
France	Private GPs provide ambulatory care and house calls. Over 2000 health centres with salaried doctors provide services mainly for the poorer segment of the population. Indirect provision (contracted)	Total freedom for people to choose and use private or public health services.	GP have no gatekeeping role and there is no referral system.	Patients pay first and are later reimbursed by their insurance. Free periodic health checks.	Patients must pay 25% of the bill unless they are covered by a voluntary complementary insurance or have a severe illness.	GPs are paid on a feefor-service basis; Salaried doctors; Salary in health centers.
Germany	Independent GPs and specialists (who are allowed to gain access to certain number of hospital beds, and highcost technical equipment). Indirect provision (contracted)	Insured people can choose their physicians, but they have to use the services of one GP at least three months.	GP do not have a strict role as a gatekeeper, since patients can also go directly to specialists.	Free at the point of delivery (only small number of services based on fee-forservice).	From 2004 a co-payments of € 10 per quarter also apply to the first contact at a GP and when other physicians are seen without referral during the same quarter.	Primary care physicians are paid on a fee-forservice basis, where the fees are adjusted to prenegotiated regional budgets using a complex calculation according to a 'uniform value scale' on region by region basis.

Greece	Nearly all primary care providers are specialists. PHC is provided in health centres (rural areas) or hospital outpatient departments (urban areas). The latest law on primary care (February 2004) provided for gradually establishing financial and administrative autonomy for primary care centres. Direct provision (contracted)	No choice (care delivered by local panel physician).	There are very few GPs and no referral system - patients go direct to specialists or to the outpatient departments of their preferred hospital.	Free at the point of delivery (high ratio of under-the-table payments)	None	Physicians are on fulltime salaries (rural health centres) or feefor-service basis (urban centres).
Ireland	All residents are eligible for all services either without charges (Category I, medical card holders, representing 29% of the population) or with charges (Category II). The latter have coverage for public hospital services subject to some capped charges but have to make a contribution towards the cost of most other services. Qualification for a medical card giving access to free services, particularly in primary care, is largely related to income and age. Indirect provision (contracted)	People in category I register with a doctor of their choice from a list of physicians. Category II people, who pay in full, are free to choose any GP or specialist.	No Yes	GP services, prescribed drugs, medicines and appliances are then free of charge Patients pay first and are later reimbursed.	None None if lower income.	GPs are paid on a capitation basis according to the patient's age, sex and place of residence. Fee-for-service if higher income; capitation (age diffentiated fee) if lower income.
Italy	Primary care is dominated by GPs, but patients often go to hospital emergency departments to avoid writing lists and prescription charges. Indirect provision (contracted)	Free choice - patients are registered with a GP.	GP acts as a gatekeeper to specialist services.	GP services are free at the point of use.	None	GP make contracts with the appropriate health unit (total 197) and are paid on a capitation basis (capitation age differentiated fees). Reforms aimed at providing additional incentives for efficiency: income can be complemented by fees for specific treatments and financial rewards for effective cost containment.

Luxembourg	Primary health care is provided almost exclusively by general practitioners. Indirect provision (contracted)	Total freedom of choice to consult any GP or specialist (high competition).	GPs have no gatekeeping role and patients can go directly to a specialist, even for primary care.	Patients pay first and are later reimbursed by their sickness fund.	5%	GPs are paid on a fee-for-service basis.
Netherlands	General practitioners play a key role since they provide most of the primary medical care. GPs are accessible 24 hours a day and they arrange out-of-hours services among themselves. Indirect provision (contracted)	Sickness fund patients must register with a GP contracted to fund, and cannot change or register with another GP inside a year.	GPs act as gatekeepers to specialist services.	GP provides free care to the patient (exception are adults with high risks).	20%, maximum € 92 annually. None if lower income.	General practitioners (GPs) are paid on a capitation basis (age differentiated fee) ifor patients insured by sickness funds and on a fee-for-service basis for the privately insured.
Portugal	Primary care is provided by integrated primary health centres (some 74% of all physicians in the health centres are GPs), extensions or health posts. GPs have a dominant role. Direct provision (employed)	People are free to choose their doctor and GPs have lists of at least 1,500 patients.	Formally, GPs have gatekeeping function to secondary care, but people often use hospital emergency departments to gain access to their preferred option of hospital care.	Free at the point of delivery.	Various None	GPs are state-employed and salaried. Private practice and additional payments such as overtime constitute significant additional sources of income. An experimental payment system for groups of general practitioners (GPs) and family doctors based on capitation and professional performance was introduced in 1999 and is under revision.

Spain	All doctors working in a given geographical area use the same primary health care centres or polyclinics. Private sector doctors have their own practices. In Catalonia, for example, primary care teamswere created at the beginning of the reform period in 1986. Sixteen years later, in 2002, a public body (the Institute of Health of Catalonia) is the major provider of these services, responsible for the management of such in 78% of cases, while in 22% of cases, primary care teams are managed by other providers Direct provision (employed)	People may choose a GP among those working in the area where they, as users, are registered.	GPs act as gatekeepers to the rest of the public health care system.	Care is free at the point of delivery.	None	GPs are mainly salaried. Doctors have their own practises and are paid on a fee-for-service basis not covered by National Health System.
Sweden	Primary care is organized around local health centres (around 950) staffed with GPs, nurses. In some urban county councils, up to 60% of primary care physicians may be private, whereas in other county councils only a few private practitioners can be found. Direct provision (employed)	Patients have freedom to choose their GP and to change their GP at any time.	None	Partially	In 2004, the fee for consulting a primary care physician varies across different county councils, and ranges from €11 to €17 Persons younger than 20: none.	Most county councils have decentralized a great deal of the financial responsibility to health care districts through global budgets. A small group of about five county councils continues to develop capitation models for primary care. The majority of GPs are mainly salaried employees.
United Kingdom	GPs in group practices (with an average of three per practice) provide primary care. There are also a small number of NHS walk-in clinics. Indirect provision (contracted)	People are free to choose GP provided that they are residents of the designated practice area. All GPs have to produce a leaflet advertising the service they provide in order to help patients choose a practice.		Free care at the point of delivery.	None	GPs are primarily remunerated by capitation according to the number of patients on their lists. On 1 April 2004, remuneration of their services moved from a system mainly based on capitation and fixed allowances to one that combines capitation and quality points.

Tab. 5 Practising Physicians (Density); Physicians per 1000 Population; (head count), 1990 – 2003.

	1990	1995	1996	1997	1998	1999	2000	2001	2002	2003
Austria	2,2	2,7	2,8	2,9	3,0	3,0	3,1	3,2	3,3	3,4
Belgium	3,3	3,5	3,6	3,7	3,7	3,8	3,9	n.a.	3,9	n.a.
Czech Republic	2,7	3,0	3,0	3,1	3,0	3,1	3,4	3,4	3,5	3,5
Denmark	3,1	3,3	3,3	3,3	3,3	3,4	3,4	3,4	2,9	n.a.
Finland	2,4	2,8	2,8	3,0	3,0	3,1	3,1	3,1	2,6	2,6
France	3,1	3,2	3,2	3,3	3,3	3,3	3,3	3,3	3,3	3,4
Germany	n.a.	3,1	3,1	3,1	3,2	3,2	3,3	3,3	3,3	3,4
Greece	3,4	3,9	4,0	4,1	4,3	4,4	4,5	4,4	n.a.	n.a.
Hungary	2,9	3,0	3,0	3,1	3,1	3,1	n.a.	2,9	3,2	3,2
Ireland	n.a.	2,1	2,1	2,1	2,2	2,3	2,2	2,4	2,4	2,6
Italy	**n.a.**	**3,9**	**4,1**	**4,0**	**4,1**	**4,2**	**4,1**	**4,3**	**4,4**	**4,1**
Luxembourg	2,0	2,2	2,3	2,4	2,4	2,5	2,5	2,5	2,6	2,7
Netherlands	2,5	n.a.	n.a.	n.a.	2,9	3,1	3,2	3,3	3,1	3,1
Poland	2,1	2,3	2,4	2,4	2,3	2,3	2,2	2,2	2,3	2,5
Portugal	2,8	2,9	3,0	3,0	3,1	3,1	3,2	3,2	3,3	3,3
Slovak Republic	n.a.	2,6	2,2	2,1	n.a.	3,6	3,7	3,6	3,1	3,1
Spain	n.a.	2,5	2,9	2,9	2,9	3,0	3,3	3,1	2,9	3,2
Sweden	2,9	2,8	2,8	2,8	2,8	2,9	3,0	n.a.	3,3	n.a.
UK	1,5	1,8	1,8	1,9	1,9	2,0	2,0	n.a.	2,1	2,2
Norway	n.a.	2,8	2,8	2,5	2,7	2,8	2,9	3,0	3,1	3,1
Switzerland	3,0	3,2	3,2	3,3	3,3	3,4	3,5	3,5	3,6	n.a.
Turkey	0,9	1,1	1,1	1,2	1,2	1,2	1,3	1,3	n.a.	n.a.
Australia	2,2	2,5	2,5	2,5	2,5	2,5	2,4	2,5	2,5	n.a.
Canada	2,1	2,1	2,1	2,1	2,1	2,1	2,1	2,1	2,1	2,1
Japan	1,7	n.a.	1,8	n.a.	1,9	n.a.	1,9	n.a.	2,0	n.a.
New Zealand	1,9	2,0	2,0	2,2	2,2	2,2	2,2	n.a.	2,1	2,2
United States	2,4	2,6	2,6	2,7	2,7	2,7	n.a.	n.a.	2,3	n.a.

Source: OECD Health Data 2005 (Database). Countries other than European are also included.

Tab. 6 Figures associated with supply of general practitioners in Italy (Source: HFA 2005 database)

	% of general practitioners with a list size >15,000 patients (GPs>15,000 Patients)/GPs*100	Patient list size (average) Registered patients/General practitioners	Density of general practitioners (GPs/Population)*10,000	Population per doctors Population/GPs
Piemonte	12.22	1119	8.2	1219
Valle d'Aosta	7.84	1058	8.4	1191
Lombardia	22.48	1176	7.66	1305
Trentino A.A.	**37.26**	**1344**	**6.57**	**1523**
Veneto	17.34	1155	7.68	1302
Friuli V.G.	10.48	1053	8.63	1159
Liguria	**5.12**	1025	8.93	1120
Emilia Romagna	18.53	1121	8.06	1240
Toscana	14.29	1038	8.72	1147
Umbria	12.62	1029	8.76	1141
Marche	12.69	1082	8.22	1216
Lazio	7.89	**994**	**9.37**	**1067**
Abruzzo	14.33	1059	8.46	1183
Molise	6.03	1026	8.77	1140
Campania	22.13	1155	7.77	1286
Puglia	16.19	1080	8.21	1218
Basilicata	9.63	1053	8.53	1173
Calabria	11.57	1063	8.48	1180
Sicilia	15.13	1080	8.16	1225
Sardegna	11.34	1033	7.85	1275
ITALIA	**15.76**	**1099**	**8.18**	**1223**

Box 1.1 The main reforms in the Italian NHS and the provision of LEA

First reform The Italian National Health Service (Servizio Sanitario Nazionale – SSN) was established in 1978 to replace the Bismarckian social insurance health care system which was established in 1943 (Law 833/78). The SSN was explicitly modeled on the British National Health Service with the declared goal of providing uniform and comprehensive care to all citizens irrespective of age, social condition, or income. Currently, the Italian National Health Service absorbs 76.4% of total health care expenditure and it is financed by general taxation (OECD health data, 2005). The fundamental imperfection in the 1978 design was that virtually the entire responsibility for financing the SSN lay with the central government, which, however, had limited power over how the USLs administered these funds. The chronic regional deficits accumulated over years reflected two tendencies of central government policy: to systematically underestimate the funding needs of the SSN and to overestimate the savings to be obtained from expenditure containment strategies.

Second reform Managerialism and regionalization were responses to the failures of the 1978 reform. National legislation from 1992 to 1993 and subsequent reforms in 1997 and 2000 have radically transformed the NHS, giving the 20 regions political, administrative, and financial responsibility regarding the provision of health care. The prime goals of the 1992 reform were macroeconomic stabilization and microeconomic efficiency. The instruments used were of two types: the health care powers and responsibilities of sub-central governments were reengineered, and policies were adopted to promote managerialism and competition. The 1992 seemed to contain the basis for transforming the highly vertically integrated SSN into a system of quasi-markets similar to that introduced in the UK in 1991.

The regions have significant autonomy on the revenue side of the regional health budget but they and are required to fund any deficit that might occur from their own resources. The distribution of funds to the regions has been based on a per capita allocation, which takes into account the regional population age distribution, mortality rates, and indicators of consumption of health care services. On the provision side the regions must deliver uniform levels of care, while on the funding side regions are mandated to cover any deficit required to provide the uniform levels of care and to use their own resources to provide services above those guaranteed by national laws. The reform laid out a new logical framework, but it took time to establish the LEAs. Economies of scale were pursued by reducing the number of ASLs from 659 in 1992 to 197 in 2000. A key innovation was the post of chief executive officer of the ASL and AO (Aziende Opsedaliere), appointed with a private performance-based contract and accountable to the region.

The Third Reform The 1999 health care reform did not aim at cost containment. Its explicit purpose was to reaffirm the original goals of universalism, comprehensiveness, and public funding of the SSN. The original goals of the SSN were to be pursued using new policy instruments. The 1999 reform promoted clinical governance but also tried to impose tighter control over the medical profession. The 1999 reform introduced significant changes in primary health care services by reinforcing group practice, introducing economic incentives for general practitioners and promoting integration between primary care physicians and district services such as social care, home care, health education and environmental health.

continued

The Fourth reform In 2001 an amendment to the constitution consolidated the power of the regions, which were given legislative power for many matters, concurrently with the state.

The Italian Constitution, revised in 2001, reserves to the central government the exclusive power to set the "essential levels of care" which must be guaranteed to all residents. Regions have virtually exclusive powers over regulation, organization, administration, and funding of publicly financed health care. The Italian Constitution of 1948 specifies the citizen's right to health. This constitutional guarantee is expressed in very general terms. Article 32 of the Constitution says that "The Republic protects health as a fundamental right of the individual and as a concern of collectivity and guar an tees free care to the indigent." The principle of a package of benefits available to all citizens irrespective of age, social condition, or income was stated later, in the law introducing the NHS in 1978. The expression "levels of care" was mentioned for the first time with the objective to guarantee equal health care coverage through out the country: "the State is to set objectives for eliminating geographical differences in social and health care conditions" (Art. 2) and "and to deter mine levels of care to be guar an teed to all citizens" (Art. 3). The same legislation also introduced another major feature of the NHS: the patient's right to choose "provider and place of treatment."

This progress was made with the agreement between the regions and the central government on 8 August 2001 which was followed by a governmental decree (the LEA decree). At present this decree is the fundamental element of the Italian health benefit catalogue. It defines the major areas of health care services to be guaranteed by the NHS (positive list), those completely excluded by public coverage (negative list), and those partially covered (only available for specific clinical conditions). The positive list is based on the recognition and systematization of current legislation (other decrees, laws, guidelines, etc.) i.e., it includes all the services that the INHS is actually providing catgorized in three macro levels of care: 1. public health services, 2. community care, and 3. hospital care. The decree also defined a system for monitoring LEA implementation across the country. Responsibility for this was assigned to a special technical body established in April 2002 and composed of representatives of the Ministry of Health, the Treasury, and the regional governments. The main objective of the commission is to "monitor and evaluate the actual provision of services included in the LEAs and their costs." In 2004 a new technical body (the Nation al LEA Commission) was established to up date LEAs on the basis of scientific, technological, and economic evidence (Ministry decree of 25 February 2005). The Commission is set up of 14 members: 6 experts of health care management, planning, and organization al sciences are nominated by the Ministry of Health, 7 are region al reprepresentatives, and one is appointed by the Treasury.

In 2001, a catalogue of SSN benefits, the LEAs, was defined in terms of a positive list and a negative list based on criteria of effectiveness, appropriateness and efficiency in delivery, and on the 'rule of rescue'. The positive list contains the services which the SSN is required to provide uniformly in all regions. This obligation is spelled out in varying degrees of detail, ranging from specific procedures to broad categories of services. Regions are free to provide non-LEA services to their residents, but must finance these with own source revenues and some actually do so. The negative list includes three categories of ambulatory and hospital services. First, a small number of services are excluded outright from SSN coverage because of their proven clinical ineffectiveness or because they are considered not to fall within the remit of the SSN. These include cosmetic surgery (except in cases of malformation and injury), ritual circumcision, non-conventional medicine, vaccinations for employment and vacation purposes and over 20 types of physio-therapy. Second, certain diagnostic and therapeutic ambulatory services, for example bone density testing, excimer laser surgery and orthodontic services, are included in the entitlement on a case-by-case basis (that is, if judged appropriate for a patient's particular clinical condition). Third, there is an indicative list of potentially inappropriate hospital admissions, classified in terms of diagnosis-related groups (DRGs), for which the regions are supposed to provide substitute treatment such as day cases and ambulatory care. Examples of these are carpal tunnel release, cataract surgery, and hypertension (Fattore and Torbica, 2004)

CHAPTER II

QUALITY INCENTIVES IN PRIMARY CARE – EMPIRICAL EVIDENCE

2.1 Introduction

This chapter deals with the empirical literature on both the application of quality indicators to assess the performance in primary care and the physician quality incentives. The aim of this review is to understand to which extent the developed indicators and the theoretical predictions on quality-related physician behaviour discussed in the previous chapter hold in practice. We start by analysing the evidence provided by several authors on both the use of ambulatory care sensitive conditions and the most direct measure of clinical quality – as provided by process indicators. Then, we broadly highlight the methodological issues to be addressed when critically appraising the evidence on physician behaviour in presence of inducement. The evidence provided is derived from several studies carried out in different settings. The empirical literature on physician induced demand is analysed from different approaches traditionally used to test for Physician Induced Demand (PID) over the past three decades. Most of these studies concentrate on physician responses to changes in doctors/population ratios (availability effect) and to fee changes. As for the latter, we specifically address the question on what the empirical evidence tell us about provider's responses to financial incentives.

2.2 Evidence on hospitalisation for Ambulatory Care Sensitive Conditions as a performance indicators

ACSCs have been widely used in different setting for measuring access in primary care. They have been employed mostly as performance indicators by health care systems. Most of the evidence is derived from studies carried out in New Zealand, the US, Spain, England, Australia and Canada and refer mainly to specific subgroups of population or conditions.

Brown and Barnett (1992) studied the influence of bed supply and health care organisation on regional and local patterns of diabetes related hospitalization. Diabetes discharge rates were found highly correlated with hospital bed supply in 5 of 8 years studied (1979-1986) confirming Roemer's Law that diabetes hospitalization depend mostly on the availability of medical care than on population need.

Jackson and Tobias (2001) conducted a study in New Zealand in which they describe the pattern in potentially avoidable hospitalizations (ambulatory-sensitive hospitalizations ASH, preventable hospitalizations PH, and injury hospitalization IH) over the past 10 years, including variations between groups differentiated by age, gender, ethnicity and degree of deprivation. They found that supply of services and access to those services also play a role. Injury hospitalization rates have increased in line with unavoidable hospitalizations, preventable hospitalization rates have declined by 40%, and ambulatory sensitive hospitalizations (the largest group of PHA) have increased by 25%. An upper age limit of 75 was used because of the prevalence of co-morbility increases for people older than 75 years. The increase in ASH is likely to have multiple causes, including: changes in incentive structures and practice patterns emanating from the health reforms of 1992-1993 (for example, altered incentive for general practitioners to refer and hospital to admit); improvements in healthcare interventions and technology; increases in the incidence and/or prevalence of some chronic diseases; increased barriers to accessing primary health care for

some population groups; and artefact arising from coding changes. This implies some caution when interpreting/analysing ACSC admission rates.

Johnston and Lynn (2004) found that avoidable hospitalization rate for children living in the least deprived areas growths faster compared to the one for children living in the most deprived areas. Also the increase in ACSH after 1st July 1997 – the year when primary care become free for children aged less than six years - was higher for children aged 1-5 years. This seems to imply that better access to primary care did not decrease ambulatory care-sensitive hospitalizations for children in the short term (although it may conceivably do so in the longer term)[29] [30].

Billings and Hasselblad (1989) found differences in hospital admission patterns between the under age 65 and 65+ population in New York City. In particular, for ACSC the age groups of 20-34 and 35-49 show the highest level of variation and the strongest degree of correlation. According to the authors, this finding can be due to four factors. The first explanation is related to the insurance that the 65+ population receives by the government. Medicare covers a considerable quantity of outpatient medical care expenses and giving an incentive to physicians to provide care through a suitable reimbursement. Secondly, intangible barriers to access to outpatient care (such as lack of a regular source of care, language difficulties, cultural factors, education, lifestyle, etc.) are supposed to be less important in older people with more experience than in younger people. Thirdly, the health status among the elderly tend to be more compromised compared to one of a young individual especially for co-morbility factors and therefore hospitalization could become unavoidable. Finally, there may exist significant

[29] They explained these contrasting results suggesting that trends in other factors might have obscured the underlying relationship between socio-economic conditions and avoidable hospitalizations for different levels of deprivation and between access to primary care and ambulatory care-sensitive analysis.

[30] In Valencia (Spain) no association was found between socioeconomic status, type of physician or a previous visit to primary care servisces and the risk of hospitalization for children due to ambulatory care-sensitive conditions once age and gender were controlled for (Casanova et al., 1996). By contrast, a comparison between the US and Spain - characterized by greater and more equitable access compared to the American ACSC one - showed that in the US the lower classes individuals make fewer visits to the doctor, they are more likely to lack indicated preventive services, they have longer intervals follow-up visits for chronic conditions and they have poorer access to needed inpatient services (Casanova and Starfield, 1995). Dafny et al. (2005) conducted a study to assess the net impact of Medicaid expansions on child hospitalizations during 1983-1996 period. The main finding is that the access effect outweighs any efficiency effect produced by expanded coverage.

economic and ethnic differences among age cohorts in some zip codes. When investigating the effect of race/ethnicity on ACS admission rates, they found some differences in disease prevalence among racial groups for some conditions. Race/ethnicity was also suggested as an indirect or proxy indicator for other characteristics (such as cultural and sociological factors) that are otherwise difficult to document.

Weissman et al. (1992) found that uninsured and Medicaid patients under the age of 65 years in Massachusetts and Maryland are more likely to be admitted to the hospital for chronic medical conditions than privately insured patients under the age of 65 years[31].

Billings et al. (1993) investigated the effect of race, disease prevalence, patient lifestyle (alcohol/substance abuse) and differences in physician decision making in New York City on hospitalization rates among different age cohorts. The largest differences between low- and high-income populations were observed in the young adult and middle-aged populations. These are the groups most likely to be affected by access problems. Individuals aged 65 years and older showed much less variation in preventable hospitalization across income strata.

The study by Begley et al. (1994) for Galveston County, Texas support the finding that individuals living in low-income areas were more likely than individuals in high-income areas to be hospitalized for some chronic medical conditions. Comparisons are made with the results documented for Maryland and Massachusetts and New York City.

Bindman et al. (1995) investigated also the role of non-access-related factors (i.e. condition prevalence, health care-seeking behaviour and physician practice style) in addition to access-related barriers (i.e. proportion of uninsured, Medicaid insured patients and proportion reported a regular place of care) played in explaining the relationship between income and preventable hospitalization rates. The analysis was limited to urban areas in California. They found that hospital

[31] They considered the following conditions: ruptured appendix, asthma, cellulites, congestive hearth failure, diabetes, gangrene, hypokalemia, immunizable conditions, malignant hypertension, pneumonia, pyelonephritis, perforated or bleeding ulcer.

admission rates for chronic diseases (asthma, hypertension, congestive hearth failure, chronic obstructive pulmonary disease and diabetes) vary with prevalence, with indicators of perceived problems, with access to medical care, and with the proportion of populations reporting no regular sources of care. Physician practice style and health care seeking beliefs did not have independent effects on preventable hospitalizations. They found a strong relationship between access and preventable hospitalization rates. Therefore, they suggested preventable hospitalization rates as a useful index of access to medical care.

Lambrew, Carey and Billings (1992) investigated the effect of having a regular source of care (considered both a measure of access and a determinant of access) on the risk of hospitalization or ACSC for a national population sample. They found that the simple presence of a regular source of care may not be sufficient to prevent such avoidable hospitalizations[32]. College education and Hispanic ethnic status appeared to be protective against hospitalization for ACS. Public insurance (Medicare and Medicaid) and presence of a chronic condition were substantial risk factors for hospitalization for these conditions. ACS hospitalizations were more likely to occur in those on Medicaid or Medicare. Yet, according the study conducted by Hayward (1991) regular source of ambulatory care was not an accurate indicator of access to health services.

Pappas et al., (1997) calculated national rates of hospitalization for avoidable conditions in the US health system by age, race, median income of zip code, insurance status. They found that rates of potentially avoidable hospitalizations were higher for person living in middle- and low-income areas, and were higher among Blacks than among Whites even among the privately insured.

Laditka et al. (2005) investigated the relationship between physician supply and hospitalization for ACSC in a large sample of urban counties of the United States. They found that in the urban areas for ages 0–17 physician supply has the largest negative adjusted relationship with ACSH. For ages 18–39 and 40–64, physician

[32] Since they found that individuals with chronic illness are more likely to have a regular source of care and since most patients hospitalized with ACS conditions have chronic illnesses, chronic illnesses is a confounder factor for the relationship between regular sources of care and ASC. Therefore, when chronic condition is controlled for in regression analysis, regular source of care is still not significant.

supply has the second largest negative adjusted relationship with ACSH. Physician supply was not associated with ACSH in rural areas. Four previous studies have examined the relationship between physician supply and ACSH with conflicting results. Two found no association (Krakauer et al. 1996; Ricketts et al. 2001), one a positive association (Schreiber and Zielinski, 1997), and the fourth the expected negative association (Parchman and Culler, 1994). A fifth study examined effects of quartile measures of physician supply using data representative of U.S. urban areas, finding that older individuals in low-supply areas had higher ACSH risk, and that those in areas having adequate supply had significantly lower risk (Laditka, 2004).

In England variations in hospital admission rates have been investigated for specific subgroups such as patients with asthma or diabetes and children. Durojaiye and Hutchison (1989) conducted in Nottingham Health District documented that improvements in primary care have not been accompanied by a fall in hospital admissions. Griffiths et al. (1996) and Aveyard (1997) documented for east London and Warwickshire, respectively, that practice with higher prescribing ratios are characterized by lower admission rates to hospital for patients aged 5-64 years and all age groups, respectively. Aveyard (1997) did not find association between staff time and hospital admission rates.

Reid et al. (1999) found at practice level that patient factors - the proportion chronically ill, the proportion unskilled (both positively related to admission rates) and the proportion who moved house in the past year (negatively related) - were the most important in explaining the variation in both overall, elective and emergency admission rates in a London health authority. Indeed, this study was not dealing with ACSCs. However, it was found (surprisingly) that variables, such as cervical screening uptake rates, minor surgery offered, and child health surveillance (i.e. proxies for quality) were positively correlated with both emergency and elective admission rates. Deprivation may affect admission rates directly through increased morbidity or indirectly through later presentation resulting in more acute symptoms or by lack of social support at home forcing

admissions. Almost 10% of the variation in admission rates was explained by the use of different local general hospitals. This can be related to an artefact, to different admission policies or to a further area deprivation effect. Contrary to common held belief, emergency admission rates were not higher for fundholders. These findings confirm the fact that hospital admission rates must routinely be adjusted for differences in patients population and hospitals used when comparing different general practices.

Giuffrida et al. (1999) found that the ranking of health authorities was affected when they were ranked by crude rates, by rates adjusted for age and sex, by rates adjusted for age, sex, socioeconomic factors and limiting long term illness reported in the census and by rates which are also adjusted for factors related to supply of secondary care. They considered three ACSCs conditions, namely asthma, diabetes and epilepsy proposed as primary care performance indicators in the UK. The regression analyses showed that a high proportion of the variance in age and sex standardised admission rates could be explained by socioeconomic and secondary care factors. Overall, these variable explained 45% of the variance in admission rates for asthma, 33% for diabetes, and 55% for epilepsy. Then, they calculated the predicted admission rates for each condition for each geographical area employing the estimates from the regression. The first was the rate predicted by using only the health variables. The second predicted rate used the health variables and the socioeconomic variables; and the third predicted rate used all the variables in the full model, namely health, socioeconomic factors, and supply of secondary care. The differences between the actual rate for an area and the predicted rates were measures of the possible effect of quality of primary care on admissions after control for possible confounding by health, socioeconomic characteristics, and supply of secondary care

Both these last studies indicated that crude admission rates are probably poor indicators of quality in primary care. Their limitations suggest the need for more appropriate measures (Jankowski, 1999).

A strand of the research has been concentrating also on the implications for quality of different payment mechanisms, such as fee-for-service, capitation, and salary. In a preliminary analysis, Dusheiko and Gravelle (2005) found evidence of a significant increase in ACS admission rates after practices switch to a Personal Medical Service contract from a General Medical Service contract. Longer the practice has been PMS more increase was observed. Non-ACSC admissions are also increased when practices become PMS and increase more the longer the practice has been PMS.

In conclusion, there are several factors to take into account when using ACSC hospitalization rates for assessing quality. These factors can be summarized as those related to: health system and use characteristics (hospital policies, physician supply, insurance system, incentive structure); demographic characteristics (age, race and ethnicity, culture and languages, population density); social and economic characteristics (education, employment status, income); population health factors (disease prevalence, severity, disability); environmental context (air pollution, poor housing, unhealthful working conditions) (Laditka et al, 2005). In particular, the evidence from several studies showed that ACSC admissions are higher in low-income areas and in areas with higher concentrations of racial and ethnic minorities and that the relationship between ACSC admissions and socioeconomic class persists even among insured populations (Billings, Anderson, and Newman 1996). This evidence suggested that other barriers to ambulatory care may exist such as transportation, inability to make child care arrangements, or lack of knowledge about how and when to engage the system with a health problem. The most common explanation for the presence of higher admission rates in low-income areas was the shortage of primary care service capacity. However, variations in disease prevalence, variations in health care seeking behaviour, lifestyle behaviours, and physician practice patterns can also affect the level of ACSC admissions. The presence of such factors does not make invalid these indicators, as long as higher ACSC admissions are interpreted not merely as unavailability of care but a sign of a more general problem in the system. Such a problem can be related to a deficiency in the accessibility of care because of

access barriers; or to shortage in the appropriateness of care (failure to emphasize prevention and wellness education contributing to the non-compliance of patients with treatment regimens, preventive health protocols, or principle of good personal health behaviour; or inadequate training and clinical standards that result in variations in physicians practice pattern); or to some environmental factors such as air pollution, poor housing, unhealthy working conditions. Unless these factors can be controlled for in study designs, they will continue to compromise the validity of ACSC as a measure of primary care system performance.

2.3 Evidence on indicators based on clinical audit

There is evidence that quality of care varies for both clinical care and assessments by patients of access and interpersonal care (Campbell et al., 2001; see Appendix LR3). The study conducted by Campbell et al., (2001) found evidence for the association between routine booking interval for consultations and the quality of management of chronic diseases; between practice size and quality of care (smaller practices scored better than larger ones for access to care while larger practices scored better than smaller ones for diabetes; more deprived practices had lower scores for interpersonal care and overall satisfaction.); between deprivation and preventive care (practices in deprived areas had lower uptake rates for cervical cytology). Preventive care and other practice variables showed no significant independent associations. Finally, team climate was found to be associated with high quality across a range of aspects of care (quality of care for diabetes, access to care, continuity of care, and overall satisfaction).

Another study by Campbell et al. (2002) assessed the quality of clinical care for angina, asthma and type 2 diabetes from medical records using the clinical audit review criteria (15 indicators for coronary heart disease, 13 indicators for asthma, 22 indicators for diabetes). Their longitudinal observational study documented improvements in the quality of clinical care in English general practice between 1998 and 2003. Patient-level quality scores were computed for all patients in 1998 and 2003. Scores were based on a simple ratio of the number of indicators for which care was actually provided divided by the number of indicators for which

care should have been provided. Expressed as a percentage, the score represents the percentage of 'necessary care' that was actually provided to each patient, within arrange from 0 to 100 where 100 represents the maximal score. Practice-level quality scores were computed as the simple average of the scores for the individual patients within each practice. Scores were also computed for individual indicators where they related to an average of less than one patient per practice. Patient-level results for individual indicators were analysed using logistic regression, and patient-level quality scores using ordinary regression, with time-point as the independent variable, and practice as a cluster variable. To determine whether practices had converged or diverged since 1998 in terms of the quality of care they provide, the variance in practice-level quality scores was compared at the two points in time using Pitman's t-test for correlated variances.

To sum up, small but steady improvements were observed in English primary care. Dusheiko and Gravelle (2005) in a preliminary draft tried to assess whether PMS contracts facilitated this quality improvements through reduction of (adjusted) ACSC admissions (assuming that these were good quality indicators). They found that switching to PMS practices increased ACSC admission rates and that the increase was greater the longer the time they had been PMS. Therefore, if adjusted ACSC admissions are actually good indicators for quality, then it could be inferred that the new contract improved quality by improving access to primary care, a better access leading to diagnose health problems for which hospitalization was unavoidable. Indeed, this may explain the increase in non-ACSCs but not that one in ACSC admissions which are admissions for conditions which ought to be treatable in primary care. They suggest that a more reasonable explanation can be given by measurement error associated with becoming a PMS practice.

2.4 Methodological issues to critically appraise the evidence

Labelle et al (1994) provide a list of some relevant criticisms directed at both theoretical and empirical approaches employed to gauge the existence and the extent of the PID, including the ones on fee changes. Labelle et al. (2004) identify three main limitations: lack of rigorous theoretical model (i.e. the fact that

the results observed in PID studies are consistent with the predictions of both neoclassical and inducement models)[33], specification error in econometric (i.e. omitted variable bias, endogeneity of independent variables and under-identification of demand equation) and measurement errors (i.e. aggregation bias and unrepresentative sampling). Some of these issues are discussed below.

2.4.1 Omitted variable bias

Omitted variable bias has been the most common statistical criticism of PID studies (Phelps, 1986). The problem occurs where a study fails to account for a relevant influence on the variable measuring PID (e.g. utilisation). It often arises because of the difficulty of incorporating important influences on service use — such as quality, medical uncertainty and non-monetary costs such as waiting and travel times.

Cross-sectional competition tests have often been criticised for this bias as they have to control for numerous regional differences in health status, income and health preferences. Not controlling for these factors places excessive importance on the variables that are included in the study, especially competition, that do vary across areas.

However, some variables can never be measured. It is, for example, extremely difficult to quantify factors such as clinical uncertainty and defensive medical practices. Certainly this leaves a lot of 'noise' in the data that interferes with a clear identification of the likely magnitude and policy implications of PID. This indicates that some omitted variable bias problems can never be adequately overcome.

'Border-crossing' presents another problem for cross-sectional studies. It occurs when patients seek medical treatment outside their residential area. This phenomenon is most important for people in rural or remote areas who often need

[33] Reinhardt (1978) made this point clear. He pointed out that the increase in supply of doctors will lead to a predictable price cuts in the case of excess supply and the increase in demand utilisation is a direct effect of this price reduction. Then, the problem the empirical studies are faced is to determine how much, if any, of the increased demand following an increase in the number of doctors is actually caused by doctors themselves.

to travel to use services. Border crossing has the effect of raising usage in areas where medical services are in high supply and reducing usage where supply is low and, thereby, exaggerating perceived PID.

Several recent competition-based studies have found no evidence of PID, in spite of a positive relationship existing between competition and service use. Often, the explanation for this result was that a previously omitted variable was driving the relationship. For example, Escarce (1992) and Stano et al. (1995) took the availability effect into consideration and found this to be the relevant factor. Carlsen and Grytten (2000) showed that consumer satisfaction increased with competition and suggested higher quality of service caused the relationship.

2.4.2 Endogenous variables

An endogenous variable is one which is determined by interaction with other variables. In a complex market such as health many important variables are endogenous — for example, price, usage, competition and quality. As variables interact within a system, endogeneity makes it is difficult to isolate the relative importance of the differing variables.

This has proved especially problematic for competition studies. One issue is that competition tends to lower price and raise usage. This relationship between price and competition means it is unclear which variable is driving the results.

Another endogeneity problem facing competition studies is that the competition (doctor/population ratio) variable itself is determined by doctors' choice of location. Doctors are likely to choose to practice in an area having regard to a number of factors such as its general amenity for living and for their ability to earn a good income in that location. The income level is likely to be influenced by factors including the price of medical services and the income and health preferences of local residents. In other terms, doctors are likely to be attracted to an area that already has high demand for health care, creating a relationship between competition and usage that is not caused by inducement.

2.4.3 Data limitations

Many of the problems described above can be attributed to the absence of data or their poor quality. A major influence on the quality and availability of data is the level at which the study is conducted. Generally, a trade-off exists between the coverage of a dataset and the detail it conveys.

Larger studies at regional or national levels have often had to rely on data collected by administrative departments. One problem with such sources is that the initial reason for collecting the data was not for research or purposes (see chapter III for a more detailed analysis). This has often implied that the effect of relevant variables (such as patient health, patient socioeconomic status or doctors' experience) on usage has not been measured — because the data have not been recorded.

Another problem facing regional analysis is that aggregation of information has been required to compare the regions, or simply to manage the size of the dataset. This results in the loss of individual information and reduces the relevance of the studies.

In contrast, some recent PID studies have been based on surveys to increase the level of detail. Evidence from survey data has supported early controversies that PID does not have a uniform affect, though it is more likely to influence discretionary services. However, because these studies are expensive to run, they have often used small samples, leaving their results open to the charge of being unrepresentative.

A final, though important, problem is that data are always affected by the system from which they are collected. As noted by Freebairn (2002, p. 295):

> *The prevalence of government intervention in the pricing and supply of health and medical services means great care and caution is required in using recorded data on prices and quantities for the econometric estimation of health demand and supply functions, including testing for PID. Observations may be on either the demand or supply curve but not both, or some likely prices and quantities recorded are for disequilibrium positions inside demand and supply.*

2.4.4 Other methodological issues

Specific methodological issues concern the studies on the financial incentives derived from different remuneration systems. The measuring of the impact of the (different) method(s) of physician payment on the volume of services consumed by the patient or on the quality provided by physicians is not trivial and some cautions need to be take into account when carrying on empirical analysis. First, for results to be meaningful it is necessary that the contexts are comparable by removing all geographical or social effects and characterized by a sufficient variability of payment methods. A second problem with this type of comparison is that it is difficult to ensure that all things apart from the payment method are in fact equal: the fact that a patient chooses a physician who is paid on a per-service basis or by capitation can be correlated with some unobservable characteristics that are themselves related to the use of the health care system. Similarly, physicians can specialize by type of contract (Glied 1998), again for reasons not unrelated to their work methods and the volume of services they provide. Finally, it is difficult to be sure that the level of payment is the same from one method to the next, and that one is not measuring the combined effect of a difference in payment method and level.

2.5 The evidence on Physician Induced Demand

Although most studies are done in the fee for service systems there are also studies in the fixed price systems that offer evidence on the existence of PID. Several approaches have been employed in the literature to empirically test the inducement hypothesis.[34] Most of these have focused on the utilisation and cost implications of PID, rather than on its effect on the health status of patients. This section looks at the findings of a selection of studies covering some of these approaches and highlights some of the difficulties in testing for PID.[35] The following approaches are discussed in the following sections:

[34] See Labelle, Stoddart and Rice (1994) for a complete listing of approaches.
[35] The profile of the different approaches in this section is largely based on Monday (2002).

1. measure the effect of doctor supply on the volume of services provided and fee levels, i.e. doctor/population ratio studies or availability effect studies (see section 2.5.1);
2. measure the effect of doctor supply on doctor-initiated visits (see section 2.5.2);
3. measure the effect of fee changes, remuneration system and expenditure caps (see section 2.5.3)
4. assess the effect of medical knowledge on service utilisation (see section 2.5.4);
5. measure variations in utilisation between small areas (see section 2.5.5).
6. measure the effect of defensive medical practices and corporate medicine (see section 2.5.6)

2.5.1 Doctor/population ratio studies

This approach represents the most commonly used test for PID. It examines how the utilisation or price of medical services change in response to changes in the number of doctors (density) in an area. The hypothesis underlying the test is that, in response to an increase in the doctor/population ratio (reflecting greater competition from other doctors) applying downward pressure on their incomes, doctors will try to induce demand or raise their fees so that they preserve their incomes. Most studies using aggregate data have examined this hypothesis and some have found evidence in support of inducement, while others have not.

Now consider the evidence for the availability effect. As argued by Donaldson and Gerard (1993, ch7) one the main sources of PID is the 3rd party payment problem. It is not surprising then that most of the research on PID is done in the US where the fee for service system associated with the 3rd party payment problem is prominent. The first extensive study was by Fuchs (1978), who analyzed the changes in supply of surgeons and the demand for operations across geographical areas in US for the period 1963-70. He found extensive evidence of inducement following an increase in the number of doctors. Cromwell and Mitchell (1986) used the same methodology and found consistent results, though with reduced estimated inducement effects.

An early Australian study (Richardson 1981), based on the framework used by Fuchs and Kramer (1972) for the United States, can be used to characterise those studies providing supporting evidence of PID. The study looked at the market for GPs and specialists in 1976 — finding that inducement was greater for the latter. Specifically, a 10 per cent increase in the supply of GPs was associated with an increase in services of between 4.6 to 5.1 per cent; and a 10 per cent increase in the supply of specialists was associated with an increase in services of between 7.6 to 11.9 per cent.

The higher inducement outcome for specialists was expected, as a smaller proportion of their visits are initiated by patients and the complexity of their services is generally greater, making patients more dependent on their advice. The author acknowledged that while the impact of supply changes on time costs — notably queuing time — and service quality had not been fully addressed, the results were unlikely to be materially affected. A more recent study by Richardson and Peacock (1999), using 1996 data, and updated in Richardson (2001), found a slightly lower inducement elasticity for GP services — around 0.4.

In a review of this latest study, Freebairn (2001) argues that the type of model used by Richardson and others adopting this approach almost certainly overestimates the magnitude of demand inducement for four main reasons. First, there is the possibility of mis-specification of the demand equation, because arguably the inclusion of hospital density and state dummy variables would yield a better specification. Hospital and outpatient services are, for example, a substitute for GP services and state dummy variables could be used as broad proxies for a variety of health status and taste variables. Second, while accepting that information asymmetry applies to some patients, its extent can be overstated and it would be desirable to extend the model to explore links between patient characteristics, the nature of medical services and the extent of asymmetry. It would also be useful to explore links between doctor time per patient and patient welfare. Third, interactions between demand and supply need to be carefully investigated, including the impact of increases in doctor supply on queuing, waiting and travel times. Such changes act to lower the effective price of medical services. Finally, quality seems to be held constant — even though an increase in

doctor supply could be expected to facilitate an improvement in quality (that is, enable doctors to spend more time with patients).

An Australian study by Scott and Shiell (1997a) examined the effect of competition on the behaviour of GPs by testing the hypothesis that GPs in areas of high competition (high GP to population ratio areas) are more likely to recommend a follow-up consultation compared to GPs in an area of low competition. In trying to improve the quality of their study with respect to the previous ones the authors includes data on GP and practice characteristics and disaggregated by medical condition. The authors concluded that their results lend some support to the PID hypothesis, but only for certain medical conditions. They identified various influences on the follow-up decision of a GP, including the age of the patient, the age of the GP, the medical condition, the size of the practice and whether a diagnostic test or medication was prescribed during the initial consultation. However, they cautioned that the results were unable to provide much guidance on the extent of PID, if it is strictly defined as whether the patient would have chosen a follow-up visit if they had the same information as the GP. They also noted that the results did not capture the effect of follow-up visits on the health status of the patient, observing that (Scott and Shiell 1997a, p. 587):

> *'This of course, does not imply that such behaviour is in any way inappropriate. It is perfectly feasible that a follow-up consultation is 'appropriate' and that the effect of competition is to encourage more appropriate care. (If this was the case, then it may be that inappropriate care was being provided in areas of low competition because of high workloads.'*

In general, evidence of increases in the utilisation of medical services linked to increases in doctor numbers is insufficient to demonstrate the existence of PID. Several other factors such as technical advances in health treatments and rising expectations on the part of patients contribute to increases in the rate of usage of medical services. Hence, a simple association between increases in the supply of doctors and increases in the usage of their services does not demonstrate causality. For example, the fact that communities respond to an increase in doctor numbers

by increasing their use of medical services may simply reflect the improved availability of, and access to, doctors.

2.5.2 Studies of doctor-initiated visits

Another technique used for testing for PID is to examine the effect of changes in doctor supply on doctor-initiated visits (rather than patient-initiated visits). The PID hypothesis is that increases in doctor numbers would lead to an increase in doctor- initiated visits (for example, so that doctors can maintain their incomes).

Wilensky and Rossiter (1981) introduced the distinction between physician initiated and patient initiated visits. Their analysis is based on individual patient data for 1977 from the National Medical Care Expenditure Survey (covering health care use and payments for the US) to examine the extent of doctor-induced demand. They found that the majority of visits to doctors in that year were initiated by patients (54 per cent). However, nearly 40 per cent were doctor-initiated. They estimated that a 10% increase in physician availability produces an inducement effect of less then 1%.

From this they concluded that, while there is a role for traditional demand analysis in explaining the use of medical services, the concept of PID, defined in terms of doctor-initiated visits, is also relevant.

Two potential problems arise with the Wilensky and Rossiter methodology. First, doctor-initiated visits do not provide an adequate proxy for doctor-induced visits (that is, visits that although suggested by the doctor were not medically necessary). Second, visits per se are unlikely to be the main method by which PID occurs — what happens *during* a visit may be more important, including length of consultation and whether referrals are made for further tests and diagnosis.

The study also found that the probability of doctor-initiated visits increased with declines in the out-of-pocket price to the patient. It is also interesting to note that a similarly structured study by Tussing and Wojtowycz (1986) covering Ireland - which also has a fee-for-service system - found a stronger positive relationship between areas with high doctor numbers and the proportion of return visits.

2.5.3 Remuneration structure studies

Analysis of doctors' payments is argued by Rice (1998) as being the most accurate test of PID because it is the most direct way of measuring the impact of changes in financial incentives on service patterns. Considerable PID research has analyzed the effect of doctors' payments on utilisation based on two main approaches, namely: whether remuneration methods affect doctors' service patterns; or how fee changes affect utilisation. More generally, the empirical evidence on the effects of different payment systems on medical services in general practice is wide and it has been reviewed in several papers (Donaldson and Gerard, 1989; Scott and Hall, 1995; Gosden et al., 1999; Gosden et al, 2001). The most recent reviews found that the empirical evidence relating to the effect of all types of financial incentives is complex and that their impact depends on clinical, demographic and organisational factors. Some specific strands can be recognized in the literature. The first bulk of research studies tend to test if doctors paid on a fee-for-service basis would have higher utilisation rates compared with doctors paid by salary or capitation. The second strand tend to test if a decrease in fees would encourage a greater volume of service provision. The third strand tries to analyze the effect of caps or fixed budget on physicians behaviour. The last groups of studies try to assess the new payment systems based on performance. Overall, this studies try to assess how problems associated with micro-(in)efficiency and macro-(in)efficiency as well as quality provision can be addressed by interventions on the fee regulation side.

In the following sub-sections we try to assess to which extent the predictions made upon the impact of different payment systems on physician's behaviour are accurate in real contexts. To this purpose and by following the main strands of literature, firstly we give evidence on how changing payment system has affected physicians' behaviour in several settings (section 2.5.3.1). Secondly, we discuss how rising fees can lead to an increase in quality (section 2.5.3.2); finally, we report evidence on the impact of the price-quantity regulation and, finally, we study (2.5.3.3).

2.5.3.1 Empirical evidence on the effects of changes in payment systems on provision of medical services

2.5.3.1.1 Capitation payment versus fee-for -service

One of the most rigorous studies on changes in the remuneration system was conducted in Denmark (Krasnik et al., 1990). This was a controlled follow-up study designed to evaluate a move from pure capitation to a mixed system of capitation/fee for services introduced to solve micro-efficiency problems (see section 1.6.1). The authors compared the level of activities before and after the mixed system implementation. There are two groups of doctors used in the study. The first one is the 'index group' or treatment group who is affected by the remuneration change. It is based on a random sample of 100 GPs from 130 GPs who agree to be involved in the study out of 265 GPs in Copenhagen city. 71 GPs completed contact sheets for all three periods in study. The second one is the control group composed by 326 GPs in Copenhagen county who were not subjected to the change in the remuneration.

Data were collected at three points in time (one before and two after). Specifically, the outcome measure were the number of face-to-face and telephone consultations; renewals of prescriptions; diagnostic and curative services; specialists and hospital referrals per 1000 enlisted patients. Predictions of the analysis are the following: *a)* GPs would increase their activities, ceteris paribus, in areas where they have more discretion; *b)* referrals to hospitals and specialists would decrease (a substitution effect); *c)* in short-term doctors would over-shoot their 'target income', with a subsequent fall as they adjusted their workload (i.e. learnt about the new system).

The results show that the number of consultations (both face-to-face and by telephone) and diagnostic and curative services rise in the group of capitation physicians whose payment changed to a mixed FFS/capitation system. Although fees were also introduced for repeat prescribing, this was significantly lower in the intervention group physicians compared with the control group after 12 months (370% relative difference). Referrals to specialists and hospitals (which were not paid for by fees) were significantly lower 12 months after FFS was

introduced into the capitation group compared with the control group. Then, this study provides strong evidence that changing from capitation payment to FFS payment presented income-generating opportunities of which GPs took advantage through increasing service intensity and workload.
The use of different data-collection instruments for the intervention and control groups, as well as possible differences in physicians and patients, might explain the differences in physician behaviour post-intervention.

Also the study carried out by Davidson et al. (1992) in USA is an important one. Here 80 volunteer PCPs were randomly allocated to either the new FFS or capitation groups and compared with PCPs on lower Medicaid fees. Children aged 18 years and below were distributed as follows: new FFS (n=1015); capitation (n=764); comparison group (n=1991). The Davidson study found that the number of primary care and specialist visits in the new FFS group was greater than the number provided by capitation physicians. This provides evidence that capitation physicians may constrain the quantity of services provided in order to minimise costs to keep within the monthly capitated amount. The number of health and emergency department visits was the same in both the capitation and new FFS groups. Capitation physicians probably have the least influence over this type of care, or perhaps these visits were being used as substitutes for primary care visits. The number of hospitalisations was lower in the capitation group compared with the new FFS group, as might be expected since capitated physicians would have the incentive to minimise these costs. Despite some evidence of lower utilisation under capitation payment, health care expenditure was higher compared with the new FFS group. The Davidson study also found that the percentage of patients receiving a number of visits to their primary care physician (either in the hospital clinic or office setting) that exceeded the number recommended in the guidelines was lower for those treated by capitation than FFS physicians. However, statistical tests were not carried out. These results need to be interpreted with caution as it was not clear whether the guidelines were evidence-based and if randomisation was successful, as physician characteristics were not reported. Unit of analysis error was also present.

2.5.3.1.2 Salary payment versus fee-for-service

The studies by Kristiansen and Hjortdahl (1992), Kristiansen and Mooney (1993b), and Kristiansen and Holtedahl (1993) are three relevant works in the literature published on the effect on activity of a change from a salary-based to a fee-for-service-based system. They used a cross sectional survey of GPs that provided information on consecutive consultations and thus were able to collect data on patients. In these three studies patient populations were similar with respect to: doctor turnover, size, travel time to doctor, travel time to hospital but were not similar with regard to GP to population ratio. GPs were comparable with respect to age, sex, number of years since graduation, place of training and job satisfaction but not number who collected patient co-payments. The authors showed that, while controlling for GP and patient characteristics, GPs paid by FFS were more likely to order certain tests, provide shorter consultations and visit patients at home The effect of the payment method was weak compared with the influence of a patient's individual characteristics such as patient age and sex - proxies for patient' health status. The physician, thus, remained more motivated by health concerns than by his income.

Another relevant study comes from Hickson et al. (1987) who conducted a 9 month controlled random trial. A number of 15 pediatricians, members of the same clinic, were randomly allocated to a system of payment (salary or FFS). The study compares the volume of preventive visits produced by physicians paid on salary and those paid on FFS: with controlled observable characteristics of physicians and patients, the pediatricians paid on FFS delivered significantly more visits. This additional volume also reflects an ambiguous variation in quality: patients treated by physicians paid per service receive better follow-up, but the volume of preventive visits suggested by these physicians departs more from the recommendations of the American Academy of Pediatrics than the volume observed for salaried physicians. According to Gosden et al. (2001), this average difference between salaried physicians and those paid on FFS basis implies two different phenomena: visits in excess of the recommended number are fewer for patients of salaried physicians than for those of FFS physicians, but the

recommended visits missed by the former are more numerous than those missed by the latter. In the Hickson study, no statistically significant differences between salaried and FFS physicians in the average number of initial and follow-up visits per patient was found. Overall completed and scheduled visits, however, were lower among salaried physicians, and this was due to fewer well child visits. To investigate whether this was due to inadequate scheduling by salaried physicians or whether FFS physicians scheduled more unnecessary well-child visits, the authors studied a random sample of patient records to determine where these visits were consistent with the American Academy of Pediatrics guidelines. The percentage of recommended visits missed by patients of salaried physicians was significantly higher compared with that of FFS physicians, but the percentage of visits in excess of a recommended number was lower, thus providing support for both explanations. Emergency visits per patient and the number of enrolled patients per physician were higher for salaried physicians compared with FFS physicians. Patients of salaried physicians were less likely to see their regular physician when they attended the clinic or emergency room compared with FFS physicians. The authors adjusted the results of this study to take into account the disproportionate career interest in private practice amongst the salaried physicians and found that the results were unchanged except for the difference in the number of patients enrolled, which became not significant. Differences between salaried and FFS physicians in four dimensions of patient satisfaction were tested but only access to physician was statistically significant. This was rated as higher for salaried physicians.

2.5.3.2 Empirical evidence on the effects of changes in fees

The most well known study is by Rice (1983) who examined rates of procedures per encounter with physicians in Colorado following administered price changes. For surgical services results were consistent with income and substitution effects working in the same direction for the cross effect and in opposite directions for the own-price effect. Nguyen and Derrick (1997) studied Medicare price reductions in 1990. Overall they did not find significant volume responses (income effect balanced by substitution effect) but for around 20% of physicians

who experienced the largest price reductions there was a significant negative income effect. For this group a one percent reduction in price led to an increase in volume of about 0.4%. It seems that there is evidence of PID following fee changes, especially in the study by Rice. Nguyen and Derrick's findings do support the predictions to the extent that doctors who are hardly hit by a reduction in their income use inducement as a way to complement this reduction. Also, since the income and substitution effects cancelled each other out doesn't rule out the existence of PID.

Rice (1983) analysed the impact of fee changes in Colorado. The study investigated how a doctor's 'practice style' was affected by new fee-setting arrangements introduced in 1977. Under the new arrangement, fees were set to the state average, which lowered fees for higher paid urban GPs and raised them for others. A one per cent decrease in fees resulted in a 0.6 per cent volume increase for medical services and a 0.15 per cent increase for surgical services.

Hughes and Yule (1992) studied British GPs, measuring the impact of fee changes on behaviour from 1966–89. They found there was no systematic response in service volumes to changes in fees. A notable finding was that salaried GPs responded to fee-for-service style bonuses for public health measures — such as immunisation and pap smears — by increasing their patients use of such services.

Scott and Hall (1995) reviewed eighteen studies based on changes in GP remuneration structures and fees. In the majority of studies that found inducement (only three of the eighteen studies showed no evidence of inducement), the impact on service usage was limited. The authors also noted that, on the whole, the statistical techniques used were inadequate. Problems included the lack of a control group, unrepresentative samples and poor use of explanatory variables.

Yip (1998) and Nguyen and Derrick (1997) both studied a fee reduction by Medicare in the US in 1990. Fees were reduced, for budgetary reasons, on services deemed to be 'over-priced'. Both studies found that volume increased most for the doctors who faced the largest fee reductions. The latter study found a one per cent reduction in fees led to a 0.4 per cent increase in volume.

2.6.3.2.1 Impact of an increase in fees on quality for night visits in UK

This section illustrates one of the most rigorous studies on the effect increase in fees on quality for night visits carried out by Giuffrida and Gravelle (2001). The setting is the English NHS where out of hours primary care is provided by GPs, by Accident and Emergency (A&E) departments, ambulance services, evening nurses, social services, pharmacists and dentists. GPs and A&E departments account for the bulk of these services. Night visits is an interesting setting to explore doctor incentives mainly for three reasons. First, out of hours care by GPs is one of the few examples in the NHS of simple fee for service reimbursement. Second, patients do not face charges. Demand is perfectly inelastic, i.e. demand is not be directly affected by a change in fee. Third, GPs have the option to provide services themselves (by being on call or on a rota with other GPs), or to pass the calls on to a deputise service (paid by the GP). Up to 1990, a single fee was paid to GP practice whether or not the visits were made by a GP or by deputising services. In 1990 the NHS introduced differential payments for visits made by GPs or by deputising services. The 1990 GP contract introduced a fee of £45 for GP visits and £15 for visits made by a deputy. It also extended the hours from 23.00-7.00 to 22.00-8.00. The reason of this policy was the explicit attempt to reduce the amount of visits by deputies. The rationale was that the visits by GPs were thought to be of higher quality, because deputies would be less informed about the patient. Most GPs regard night visits as a negative aspect of their career. Night calls are seen as a source of stress that disrupts family life.

Giuffrida and Gravelle (2001) assume that GPs maximise a utility function which capture this labour-leisure trade-off. Utility depends on GP's level of income (π), on leisure (L) which in turn depends on GP visits ($L(g)$), and on the demand management activity (a) which is the opposite of supplier induced demand since doctors try to reduce demand for this type of visits.

$$u(\pi, L, a) \qquad [2.1]$$

The practice income depends on the income from doing night visits themselves (gf^{g}), the income deriving from providing visits through deputising services

($(n-g)f^d$), the total cost which depends on the total demand of visits ($c(n)$), plus other income the practice receives (y), for instance from capitation

$$\pi = gf^g + (n-g)f^d - c(n) + y) \qquad [2.2]$$

Where f^g and f^d are the fee per visits for GPs and deputising services, respectively.

Consider two extremes (**Errore. L'origine riferimento non è stata trovata.**). The first extreme is given by the case in which all visits are performed by deputising services ($g = 0$). The practice income equals $\pi = y - c(n) + f^d n$. The other extreme is when all visits are performed by the GP practice ($n = g$). In this case the practice income is $\pi = y - c(n) + f^g n$. The line between the two extremes represent a budget constraint. Doctors may prefer much more leisure or much more work. This implies staying on two indifference curves, I_1 (doctor 1) and I_2 respectively (doctor 2).

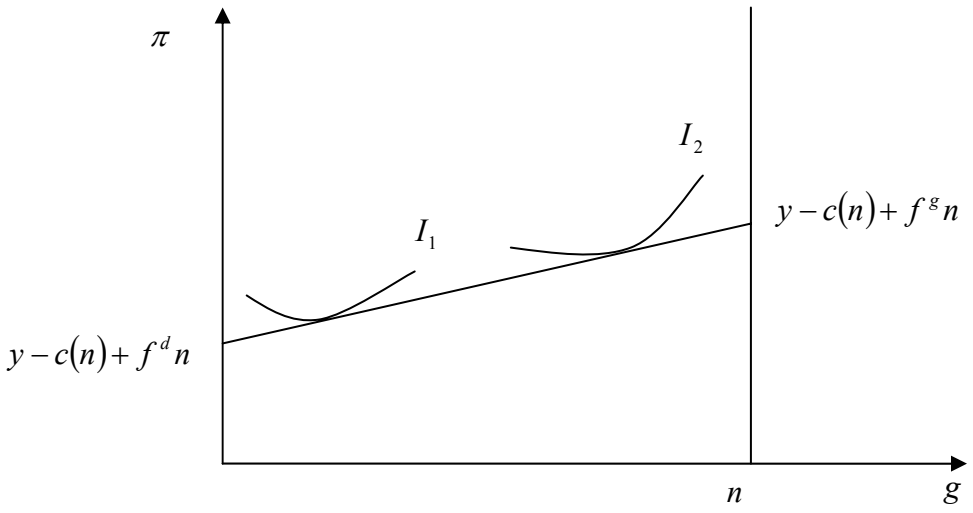


Fig. 25 Change of night visits

Now, we try to analyse the effect of change in the gap between f^d and f^g (Fig. 26). For individual 1 the income effect will be in the same direction as the substitution effect. He potentially increases the number of visits. For individual 2 budget constraint is higher and the number of visits fall down. The income effect works in the opposite direction of the substitution effect.

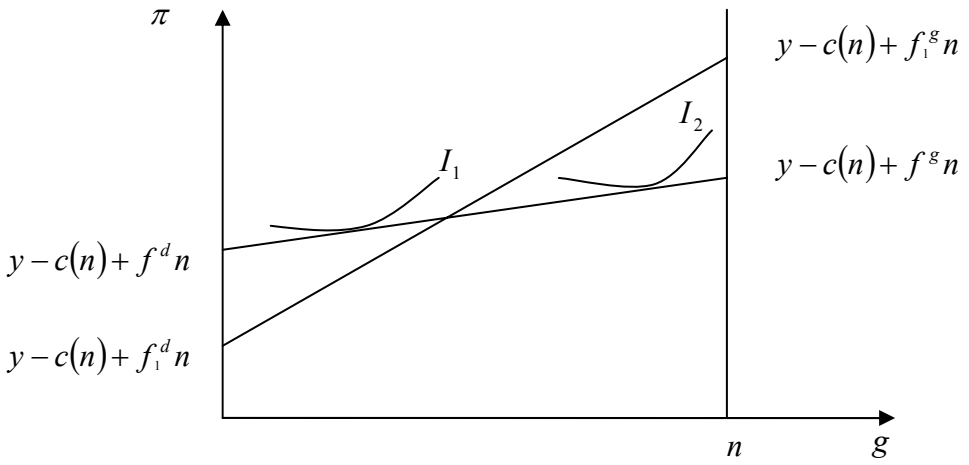


Fig. 26 Effect of changes in night visit fees

Giuffrida and Gravelle (2001) analyse also the effect of an exogenous increase in demand for night visits, from n to n_1. What happens at the budget constraint depends on the impact of the increased demand on the marginal cost. When demand goes up the level of income and fees are unchanged, the income from providing night visits goes up from $f^d n$ to $f^d n_1$. But this is offset by an increase in total cost for providing visits. This lead the budget constraint to shift down (Fig. 27).

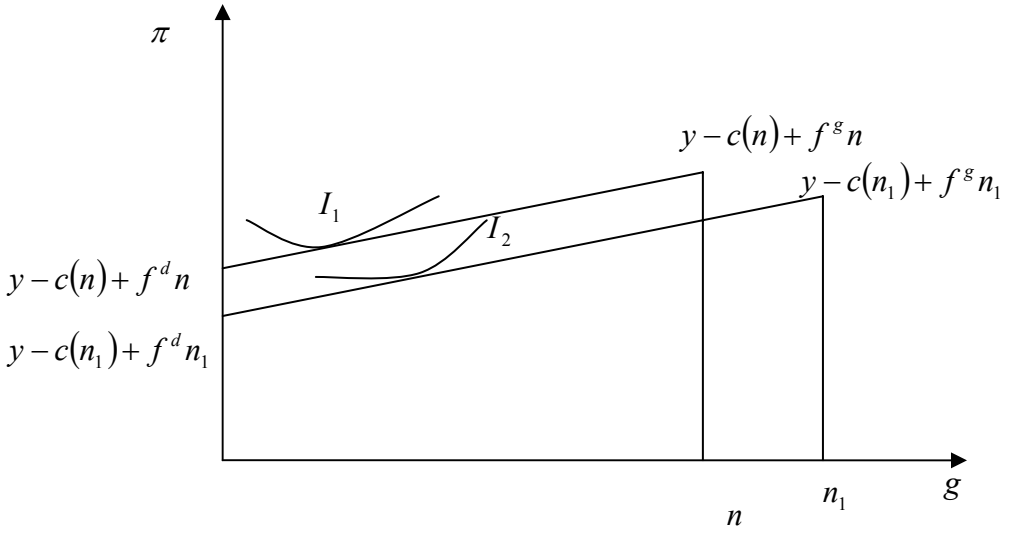


Fig. 27 Effect of increase in demand

Given this theoretical framework, Giuffrida and Gravelle (2001) use data from Health Service Indicators database supplied by the NHS Executive which cover 11 financial years. They present evidence that the number of GP night visits responds positively to an increase in fees. There is also evidence of a reduction in the use of deputising services when differential fees were introduced as part of the 1990 GP contract.

2.5.3.3 Evidence on the effect of price-quantity regulation

This section describes one of the most rigorous studies on the effect of price-quantity regulation on GPs behaviour carried out by Rochaix (1993). Her work can be thought as a 'natural experiment' performed to identify responses to exogenous shocks in the remuneration system by analysing the physician labour supply. Her framework on reimbursement can be summarised as follows. The total reimbursement a doctor receives depends on three components:

$$R = W \times S \times F \qquad [2.3]$$

The first one is the workload, i.e. the number of cases; the second one is the intensity of treatment i.e. the number of services provided per case; and the third one is the fee per item of services. In terms of regulation, it is possible to control the fee i.e. the fixed amount to pay for each particular procedure (F); or to regulate the cost for case, i.e. how much the doctor is paid for treating a particular type of patient $(S \times F)$; or regulate the overall level of income (R). Then, it is clear that regulation of F alone cannot control total expenditure.

Regulation of FFS rates

In a fee for services system, it may be possible for payers to set fee schedules (or rates of reimbursement per item of service), rather than have fees determined by competition. Given $R = W \times S \times F$, where R is the total reimbursement, it is clear that regulation of F alone is not sufficient to control total expenditure. Doctors may influence their reimbursement by modifying his workload or by increasing the intensity of treatment provided per case or by changing the mix of services shifting to more complex methods and more expensive procedures.

Regulation of cost for case

An alternative is to regulate cost per case rates ($S \times F$). This removes the ability of physician to compensate for lower fees by increasing resource use. However, total reimbursement can be still increased by increasing the number of case treated (W) unless this is also controlled. Payment is based on average costs for each diagnostic group (yardstick competition).

Regulation by capitation or by global budgeting

One way in which the incentive to enhance utilisation can be tackled directly, and in which control of total expenditure is enhanced, is by the imposition of regulated capitation rates. Reimbursement is now a function of patient affiliation, and does not vary directly with either utilisation or resource use. This is the sort of regulation inherent in Health Maintenance Organisation (HMO) arrangements in the US, and in setting the budgets for GPs in the UK. In this mechanism providers

face incentives to constraint resource use per case (as with the cost for case-regulation) and also resource use overall.

The policy context of the analysis was given by a study of Barer et al. (1992) where it was shown that the fee controls in Québec during 1970-76 were more than outweighed by substantial increases in output per doctor (i.e. changing in W and S). In response to that, in the attempt to contain overall cost to solve macro-efficiency problems in 1976 Québec adopted price-quantity regulation (see Chapter I). This was the first example of global cap on physician's earning. The government and the medical profession negotiated a 'target income', based on previous level of income, inflation and physician density forecasts. It was translated into overall rate of fee increase. The idea behind this was that if the level of activities exceeded the income target, this led to a reduction of fee increase in following year (i.e. a collective sanction). This was seen as unfair since high activity GPs penalise the rest. In response, a quarterly ceiling for high activity was introduced so as to discourage 'workaholic' physicians from billing beyond a threshold amount over a quarter. A fee reduction of 75% for GPs exceed ceiling was introduced (ceiling effect). Concurrently a fifteen month freeze of fee was imposed (fee freeze effect). Then, this was a setting where great majority of physician incomes cames from fee-for-service (94.5% in 1978, 87.8% in 1983)

The data were a panel data set of 677 Québec GPs between 1977 and 1983. Analysis concentrates on 53 procedures. Doctors were grouped into four sub-sample. The first group ($G0$) was composed by doctors with low income or semi-retired ($n = 86$); the second one ($G1$) by doctors with low activities and income between the minimum standard and the target income ($n = 360$); the third one ($G2$) by doctors with an income between the target income and the cut-off of 10% below the ceiling ($n = 119$) and the final group ($G3$) by doctors with income above the cut-off ($n = 112$).

The author examined the effects of the concurrent 15-month tariff freeze and removal of quarterly expenditure caps (Nov 1 1979 to Jan 31 1981), and the

subsequent 9-month period with only removal of expenditure caps, which coincided with a physician strike (Dec 1 1981 to Aug 30 1982).

In the model it is assumed that the GPs have to take a two-stage decision. The first decision is how much time devote to their practice and how much time to spend for leisure. The income-leisure trade-off is influenced by the real wage:

$$\ln Y = \alpha_1 + \beta_{11} July + \beta_{21} Dec + \beta_{31} NC1 + \beta_{41} NC2 + \beta_{51} D3 + \beta_{61} \ln P_A + \beta_{71} \ln CP1 + e_1 \quad [2.4]$$

Having decided on an activity rate, the second decision is how much effort allocate across activities/procedures. This is influenced by relative fees:

$$\ln Q_i = \alpha_2 + \beta_{12} July + \beta_{22} Dec + \beta_{32} NC1 + \beta_{42} NC2 + \beta_{52} D3 + \beta_{62} \ln P_i + \beta_{72} \ln P_A + \beta_{82} \ln Y + e_2 \quad [2.5]$$

where

Y - physician's income, Q - procedures carried out; $July, Dec$ - dummies for holidays periods; P_A - overall fee index, P_i - price of procedures, $CP1$ - overall price index; $NC1$ - period of tariff freeze + ceiling deregulation; $NC2$ - ceiling deregulation + strike in hospitals; $D3$ - third month in quarter

The author estimates separate equations for different procedures. She found that in response to fee freeze, doctors increased the overall level of activity in order to maintain their income. In addition, during the tariff freeze period, doctors shifted towards more complex procedures (specifically, from ordinary examinations to major medical examinations, and from these to complete medical examinations). Finally, she finds that GPs reduced medical activity in response to quarterly income caps.

2.5.3.4 Evidence on pay-for-performance system

In spite of the assertions of its proponents, the empirical foundations of pay-for-performance in health care are rather weak. There are only a few studies

demonstrating that pay-for-performance leads to improved quality of care (Amundson et al., 2003; Fairbrother et al., 1999; Kouides et al. 1998; Ritchie et al. (1992); Roski et al, 2003). A review can be found in Rosenthal and Frank (2006). Specifically, Kouides et al. (1998) conducted in a Medicare population a non-randomised controlled trial to compare target payments for increasing influenza immunization rates *vs* FFS. The two groups, thus, differed in terms of practice size and specialty mix. The intervention group included 62 physicians admitting to a single hospital, while the comparison group of 82 physicians was drawn from the same community. Primary care practices in the intervention group received an additional 10% ($0.8) or 20% ($1.6) reimbursement per flu shot according to whether they immunised 70% or 85% (respectively) of the eligible population. PCPs in the control group –paid by FFS- only received the fee for each immunisation of $8. With approximately 325 Medicare patients, the median practice in the intervention group could earn up to $560 in additional reimbursement per year. They found that the physicians receiving fees plus target payments had an influenza vaccination rate 9.4% higher than the FFS group, but this was not statistically significant. The authors also estimated that the additional cost per extra immunisation gained using the target payments incentive was $3.02.

Another study was performed by Ritchie et al. (1992). This was an interrupted time series study conducted in Scotland to examine the introduction of target payments into a FFS system. They found that immunisation rates improved after target payments replaced FFS, but using logistic regression analysis the authors found no evidence that the overall linear trend in immunisation rates had changed as a result of the target payments. In this case, primary care practices received a lower or higher payment according to whether they immunised 70% or 90% (respectively) of the eligible population.

Roski et al. (2003) studied the impact of financial incentives with and without access to a patient registry to support adherence to smoking cessation guidelines. A total of 40 clinics were randomized to a control group, a financial to a computerized patient registry linked with telephonic smoking cessation

counselling. The study examined the impact of the two interventions compared with each other and the control group on physician documentation of smoking status and advice to quit smoking (which were explicitly targeted), as well as patient smoking cessation rates after 1 year. Financial incentives were designed on the basis of reaching the following fixed performance targets: (1) documentation of smoking status for 75 percent of all patients older than 18 and (2) documentation of advice to quit smoking at the last visit for 65 percent of current smokers. Small clinics (those with fewer than seven physicians) were eligible for a $5,000 bonus; larger clinics were eligible for a $10,000 bonus. Clinics meeting one but not both targets received $2,500 or $5,000 depending on size of the practice.

In this experiment, financial incentives improved both documentation of smoking status and advice to quit compared to the control group. Compared with the control group in which identification of smokers increases by 6.2 percentage points during the 1-year study, clinics receiving the financial incentive alone improved identification by 14.1 percentage points (a statistically significant 7.9 percentage point difference). Similarly, the control group improved advice rates by 18.3 percentage points, while the financial incentive group improved by 24.2 percentage points. Despite improved adherence to guidelines, however, there was no significant impact on smoking cessation rates. In addition, in a somewhat puzzling result, the clinics that were offered both the financial incentive and access to the patient registry and telephonic counselling system showed no improvement relative to the control group.

Amundson et al. (2003) in a recent study in the Health Partners system in Minneapolis examined an intervention in which 20 medical groups varying in size from16 to 500 physicians were offered bonuses for four clinical quality measures. One of these measures captured whether the group's physicians asked patients about smoking status and counselled identified smokers to quit during office visits (so-called ask and advise rates). The eligible medical groups were responsible for between 2,939 and 141,411 Health Partners enrollees and could receive a bonus of

between $6,650 and $43,750. To receive the award, groups had to achieve 80 percent on both components of the measure (i.e., asking and advising). In addition to the financial award, groups achieving the targeted level of performance were publicly recognized. All participating groups received feedback on their performance at baseline and at 1-year intervals. At baseline, none of the 20 medical groups met the standard; the average ask rate was 49 percent, while the average advise rate was 32 percent. During a 3-year period, identification of tobacco use increased by nearly 25 percent, and advice to quit increased by more than 50 percent.

However, how much of this change was due to the financial incentives is unclear for two reasons. First, there was no control group, so secular trends over the 3-year period may account for some or all of the observed increase in compliance with tobacco cessation guidelines. Second, the impact of the quality bonus cannot be disentangled from the effect of the performance feedback that medical groups received at the same time, which some prior studies indicate may also be an effective motivator.

Fairbrother et al. (1999) studied the impact of financial incentives paired with performance feedback on childhood immunization rates in a low income urban population In this study, 60 office based paediatricians serving high proportions of Medicaid-enrolled children were randomized to one of three intervention groups or a control group. The first intervention group received feedback only on their immunization performance at three intervals of 4 months each. There were two financial incentive groups, one in which physicians were paid a performance bonus for achieving a population target immunization rate and a second that paid enhanced fees for each timely immunization. Both of these groups also received feedback on their performance. For the bonus group, awards were offered both for improvement from baseline immunization rates and absolute performance levels. Physicians were offered increasing financial bonuses for a 20 percent or 40 percent improvement from baseline ($1,000 and $2,500, respectively) and for reaching 80 percent coverage regardless of baseline performance level ($5,000).

For the enhanced fee intervention, physicians received $5 for each vaccine they administered within 30 days of its coming due and $15 for visits in which multiple vaccines were administered and all due vaccines were up to date. Neither feedback alone nor the enhanced fees improved the likelihood of childhood immunization in the study population. The rather sizable bonus did improve immunization rates, but this was primarily achieved through better documentation of immunizations children had received outside of the practice, suggesting no real gain in quality of care.

There are two limitations to the Fairbrother et al. (1999) study that should be noted. The sample size of only 15 physicians in each of the four assignment groups resulted in a study with severely limited power to detect anything but the largest effects. Equally problematic was the 8-month time horizon over which the program was studied. Had the intervention continued for a longer time, the paediatricians in the bonus group may have moved from improved documentation of immunizations received outside of their office (likely the lowest-cost way to improve their scores) to increasing actual immunization rates.

Doran et al. (2006) showed that in the first year of the pay-for-performance program, English family practitioners performed extremely well with respect to the quality targets, attaining a median of 96.7 percent of the available points for clinical indicators. This greatly exceeded the 75 percent predicted when the scheme was negotiated, and consequently the cost to the tax payers was considerably more than expected. Before the new contract was implemented, family practitioners typically earned from £70,000 to £75,000. The pay-for-performance program increased the gross income of the average family practitioner by £23,000, although the practitioners bore any additional nursing and administrative costs of meeting the targets. However, there is no baseline with which to compare performance in the first year of the U.K. program. The high levels of achievement might suggest that the targets were too easy to achieve. The scheme has been revised for 2006–2007: all minimum and some maximum payment thresholds have been raised, 30 indicators have been dropped or

modified, and 18 new indicators have been introduced. The high achievement levels might also have resulted from misreporting by practices. To counter misreporting, Primary Care Trusts, statutory bodies responsible for the delivery of health care in local areas, inspect all local practices and undertake detailed audits of randomly selected practices and those suspected of incorrect or fraudulent returns. The results of these audits are not, however, publicly available. Because achievement was universally high, there was little variation between practices. It was not surprising, therefore, that socioeconomic and demographic factors had relatively little influence on achievement. Smaller practices performed marginally better overall than large ones, although there was much greater variation in the performance of small practices, and many smaller practices are believed to have merged in the face of the administrative pressures from the new contract.

It is possible to draw some conclusion from this strand of literature. First, physicians with baseline performance already above the target seem to understand that they need only to maintain the status quo to receive the bonus payments. More surprising, perhaps, is that low-performing groups improve as much as they can, given that their short run chances of receiving a bonus were likely to be low.

Paying explicitly for quality improvement could alter the incentives for high-performing and low-performing groups, allocate bonus more toward the latter group, and possibly increase the overall impact of pay-for-performance. In addition, there could be a fairness concern whenever the low-performing groups face insuperable barriers to achieving the target because of limited resources or a patient population of low socioeconomic status. Some payers, however, disapprove to reward improvement rather than achievement because it in effect excuses low levels of performance. Paying for improvement fails to reward and even penalizes providers that have already achieved high levels of health care quality at the time a pay-for-performance program is initiated. It is possible to reward both performance and improvement and thus try to achieve multiple objectives. It is also possible that financial rewards for quality can be too low to stimulate significant departures from the underlying trend in quality improvement.

Finally, it is important to take into account the fact that substantial quality improvement could take time. To alter the underlying rate of improvement, physicians may need to make investments in infrastructure and human resources, and these investments could derive from cash flow created by bonus payments.

In many settings, new pay-for-performance initiatives represent the first time that quality-of-care data are being systematically collected and, in some cases, publicly reported, making it difficult, if not impossible, to isolate the contribution of the payment incentives.

Most current pay-for-performance programs, should be viewed as a first step in the direction of aligning payment incentives with health system quality goals. Realization of the full potential of pay-for-performance to reduce the persistent gap between evidence-based and actual practice will require that payers adapt their incentive strategies as evidence to support best practices accumulates. In addition, as Marshall and Harrison (2005) point out, payers has to take into account that there may be some dangers of unintended negative effects in terms of crowding out (see section 1.9).

The principal lesson we draw from this studies is that incentive design matters. The growing evidence from the continuing experimentation with pay-for-performance in the market will highlight these initial findings and other potential design lessons.

2.5.4 Medical knowledge and utilisation studies

A few studies have been carried out to test the proposition that PID arises from an imperfect agency relationship related to asymmetric information between doctors and their patients. According to Mooney and Ryan (1993) studies of PID have failed to provide direct tests against the benchmark of 'fully informed patients'.

Hay and Leahy (1982) performed a study to identify whether doctors and their relatives use more or fewer medical services than other patients. A finding

consistent with PID would show that doctors and their families make less use of medical services than other patients. However, the authors found that medical professionals and their families were as at least as likely to visit physicians as others after controlling for factors such as access to care, ability to pay, perceived health status and socio-economic status.

A similar study by Bunker and Brown (1974), looking at surgery rates between doctors and their spouses and non-health professionals and their spouses, yielded a similar result.

However, according to Rice (1998), there are at least two problems with such studies. One is the difficulty of accounting for the fact that doctors and their families are often able to secure medical services at more favorable prices. Another is that health professionals may demand more services in an attempt to minimize the impact of medical uncertainty on their treatment.

2.5.5 Small area variation studies

A number of studies have identified substantial variations in rates of medical procedures and surgery across small areas (that is, regions within a country or state). Some analysts maintain that unexplained variations may be attributable to demand inducement by doctors.

Richardson and Peacock (1999), for example, report the results of a comparison of variations in rates for 15 medical procedures covering statistical local areas in Victoria. They identify substantial differences between these areas when the variation predicted by the age/sex composition of each area is compared with the actual variation. For example, the actual variation exceeds its predicted variation by a minimum of 110 per cent for a total hip replacement and by 2000 per cent for a colonoscopy. In their assessment, Richardson and Peacock (1999, p. 6) note that:

> *The inescapable conclusion appears to be that the dominant factor in allocating these services is the clinical judgement of doctors. It is simply not credible that, with the removal of significant income and price*

barriers, such variation could arise from differences in individual patient preferences.

Some small area variation studies have been criticized on methodological grounds for failing to control adequately for possible explanatory variables such as differences between areas in the age, sex and medical condition of patients, as well as institutional variables such as differences in insurance coverage. Further, relatively large differences in usage rates can arise from chance alone. As noted by Diehr (1984), if usage rates are normally distributed, the highest and lowest rates will, on average, differ by 2.3 standard deviations for comparisons involving five small areas, even if the underlying rate is the same in all areas.

Within the literature, the underlying reasons for small area variations and their policy implications have generated controversy. While some see PID as a part explanator for the variations, others have suggested that a doctor's practice style (linked to their beliefs, habits and practice patterns) is likely to be a distinct and important contributor (see, for example, Folland, Goodman and Stano 2001). McPherson et al. (1981) note that broadly similar patterns of variability are observed across countries, including those where physicians have no or only limited real financial incentives to induce demand.

Nevertheless, for Richardson and Peacock, as well as others (see, for example, Wennberg 1988), these studies highlight an important point — actual medical decision making is characterized by extensive uncertainty in relation to 'appropriate' medical practice. Further, these small area variations need not imply a breakdown in the agency relationship. As noted by Feldman and Sloan (1988, p. 252):

Decisions made with imperfect information and uncertainty may characterise both patient and physician behaviour in most medical markets, even though the physician acts according to his perception of the patient's best interests.

At a policy level, small area variations have provoked debate about whether they point to large and potentially avoidable social costs due to 'inappropriate care' (

that is, over-use, under-use and misuse of medical procedures in relation to patient needs). If so, there could be scope for improving the quality as well as the efficiency of medical care. One practical development arising from this debate has been a growing interest in evidence-based medicine, including initiatives to promote evaluations of medical procedures and develop clinical guidelines.

2.5.6 Other evidence on PID: defensive medical practices and corporate medicine

McGuire (2000) refers to defensive medical practices and corporate medicine as two examples supporting PID. Some surveys indicate that doctors adopt defensive medical practices for fear of malpractice litigation. Some of these practices may involve PID-type behaviour - such as increased levels of servicing (ordering more diagnostic tests, undertaking more follow-up visits and making more referrals to specialists).

According to the survey results published in Hancock (1993) and conducted in Australia, between 38–85 per cent of doctors often or occasionally adopt defensive practices, with the incidence of these practices amongst GPs being more prevalent than specialists. Foremost amongst the practices adopted by GPs are more detailed note taking, referrals for non-invasive diagnostic procedures (such as blood tests and X-rays) and follow-up consultations (Tab. 7). Furthermore, these types of defensive practices had increased significantly over the previous five years.

Tab. 7 Proportion of doctors adopting certain practices because of the threat of litigation, 1992a *Source:* Hancock (1993)

Practices adopted	*All doctors*	*GPs*	*Specialists*
Follow-up consultation	72	76	64
Referral for a second opinion	59	65	49
Referral for invasive diagnostic procedures	52	58	40
Referral for non-invasive diagnostic procedures	72	78	62
Referral for investigative procedures with a known element of risk	38	42	31
Arranging tests which are unlikely to influence patient management	53	58	44
Avoiding high risk procedures	54	63	39
Avoiding prescribing certain drugs or devices	61	70	44
More detailed note taking	85	88	79

a Percentage of doctors reporting that they 'often' or 'occasionally' adopt these practices — the latter being the most predominant. Between 1064 and 1123 doctors responded to the survey.

There is also support for the view that some of the costs associated with defensive medical practices may be avoidable through reform to malpractice liability arrangements. For example, Kessler and McClellan (1996) concluded that reforms to malpractice arrangements in the US, which reduced the fear of litigation, would cause reductions in medical expenditures of 5–9 per cent.

The growth of corporate medicine has given rise to concerns about the possibility of conflicts of interest between doctors and patients and, as a consequence, the possibility of PID. McGuire (2000) cites three overseas studies suggesting that doctors with financial interests in diagnostic testing and therapy facilities refer patients to these facilities more often and provide more services per patient (see Hillman et al. 1990, Hillman, Welch and Pauly 1992, and Mitchell and Sass 1995). Evidence of the possible influence of ownership characteristics on some GP treatment practices in Australia is provided in a recent study published by the Australian Institute of Health and Welfare (AIHW 2001b). An analysis of X-rays ordered by Australian GPs finds a 'definite trend of increased order rates with increased size of practice'. Doctors working in the larger general practices order more diagnostic imaging tests than solo- practitioners and other small practices. Many of these large practices may be owned by the same health care corporations that own the imaging facilities. The higher rates of service where co-ownership of primary care and imaging facilities is involved may suggest some form of 'interference' with doctor referrals. That is, doctors working for a corporation may face pressure to act, in part, as an agent of the corporation (Fitzgerald 2001). However, it may also indicate other motives on the part of doctors, such as access and convenience.

2.6 Conclusions

This chapter has discussed the evidence on the mechanisms used to influence quality. A great number of studies from health economics literature dealt with this topic. Yet, many empirical studies are based on natural experiments and are therefore opportunistic. Only a few studies have been able to control accurately for the many other differences between GPs that may influence their behaviour. The most notable omissions are patient characteristics, case mix and GP characteristics. These omissions are particular relevant from a theoretical perspective.

Moreover, the bulk of research has been on the impact of payment systems on behaviour of physician in post and has ignored the implications for the medical labour market. The focus has been on the marginal incentives for different types of activity (visits, referrals, prescribing, etc.). As well as influencing clinical decisions, payment systems are major determinants of professional job satisfaction and morale, and can therefore influence supply and labour market behaviour. For example, recent payment reform in UK general practice, which has introduced more flexible payment system and a wider variety of NHS contract has been introduced to improve professional morale, job satisfaction, recruitment and retention. The argument was that the supply of doctors to general practice depended both on the level of payment and on the method. There has been very little research on the influence of different payment systems on labour market participation in general practice.

CHAPTER III

DOES BETTER CLINICAL QUALITY IN PRIMARY CARE REDUCE ADMISSIONS FOR AMBULATORY CARE SENSITIVE CONDITIONS?

Abstract

Context. Admission rates for ambulatory care sensitive conditions are often used as a performance indicator for primary care since they may reflect access to primary care and its clinical quality. The United Kingdom provides a setting with universal access to primary care which allows the relationship between clinical quality of care and admissions to be tested independent of access effects.

Objective. To determine the relationship between quality of clinical care provided in primary care practices and admissions for relevant ambulatory sensitive conditions.

Setting. Longitudinal study of 60 family practices in England: pooled cross-sectional [panel data] analyses of data from 1998 (60 practices) to 2003 (43 practices)

Design. Quality of care assessed using validated quality indicators applied to the medical records of patients with diabetes, asthma and angina (mean number of records per family practice, per condition, per year 15 - 21).

Outcome measures. Hospital admission rates for diabetes, asthma and angina.

Results. A 10% improvement in angina quality was associated with a reduction in angina admission rates of 6.3% (95% CI, -10.2% to -2.3%). For asthma there was no significant effect of quality. A 10% increase in diabetes quality was associated with a 9.5% (95% CI, 2.2%, to 16.8%) increase in diabetes admission rates. Morbidity, socio-demographic and geographical factors had more effect on admission rates than clinical quality. Angina admission rates showed modest sensitivity (63%) and specificity (57%) as a predictor of above average quality.

Conclusions. Clinical quality of care explains a small proportion of the variation in admission rates, and its effect is not always in the expected direction. Caution is required when using ambulatory care sensitive admission rates as an indicator of clinical quality in primary care.

Acknowledgements

I would like to thank Mark Dusheiko (University of York), Dr Stephen Campbell (University of Manchester), Liz Middleton, David Reeves, Prof. Martin Roland (Univesity of Manchester) and Hugh Gravelle (University of York) for their collaboration, support with providing data and helpful comments.

Preliminary versions of this paper was presented in 2006 at the 2nd conference organized by HESG (Health Economics and Research Group) in London, and at the 11th conference in Health Economics organised by AIES (Associazione italiana di Economia Sanitaria) in Venice.

3.1 Introduction

Admissions for ambulatory care sensitive conditions (ACSCs) are potentially avoidable hospital admissions where timely and effective intervention outside hospital can reduce the risk of hospitalization by 'either preventing the onset of an illness or condition, controlling an acute episodic illness or condition, or managing a chronic disease or condition' (Billing et al., 1993). Higher ACSC admission rates may be due to poorer access to primary care or to poorer clinical quality when accessed (Giuffrida et al. 1999, AHQR, 2004; 2005; NHS Executive, 1999; Niti and Ng , 2003; Jackson et al., 2001; Billings et al, 1989; Weinberger et al. 1996; Weissman et al, 1992; Bindman et al. 1995; Pappas et al. 1997; Blustein et al. 1998; Ricketts et al. 2001). Admission rates for ACSCs have been used as health system performance indicators in the US, UK, Canada and Australia to identify potential access or quality problems in primary care because of the difficulty in obtaining more direct measures of access or clinical quality (Dafny et al. 2005; Fireman et al. 2004; Brown et al. 1992; Begley et al. 1994).

There are a number of problems in using ACSC admission rates as performance indicators for primary care. Variations in ACSC admission rates may be associated with factors considered beyond the control of primary care physicians such as the health, demographic and socio-economic characteristics of patients, and the availability of hospital care ((Parchman et al 1994; Billings et al. 1996; Lambrew, 1996; Krakauer et al, 1996; Laditka et al. 2003; Millman et al. 1993; Moore et al., 1989). ACSCs are also relatively uncommon events and so variation may simply be due to chance (Lambrew, 1996). It is also important to know whether variations are due to variations in accessibility of primary care or in clinical quality when it is accessed since this will affect the appropriate policy response.

Most studies of ACSC admission rates have been concerned with access to care, NHS Executive, 1999; Jackson et al., 2001;' Billings et al, 1989; Bindman et al. 1995; Pappas et al. 1997; Bindman, 1995; Pappas, 1997; Laditka, 2003) and there

is little evidence on the extent to which better clinical quality in primary care can reduce hospitalisations for ambulatory care sensitive conditions. Few studies have examined the relationship between direct measures of the quality of clinical care derived from patient medical records using evidence based review criteria, and ACSC admission rates (AHRQ, 2005; Marshall, 2002).

This paper investigates the validity of ACSC admission rates as an indirect measure of clinical quality. We examine the relationship between ACSC admission rates and directly measured clinical quality indicators for diabetes, asthma and angina. The study was carried out in English family practice, a setting where there is universal access to primary care which therefore allows the relationship between clinical quality and admissions to be tested, independent of issues relating to whether or not the patient has a regular provider.

3.2 Methods

3.2.1 Sample

The data is based on a longitudinal observational study of 60 nationally representative family practices in six geographical areas of England collected between 1998 and 2003. The study practices were a random sample of practices stratified for size of practice, deprivation of its population, and training status of the family practitioners. The practices are described in more detail elsewhere (Campbell 2005; Gnanadesigan, 2004). They have an average of three family practitioners and 6000 registered patients each. Quality of care was measured from patient records in all 60 practices in 1998 and in 43 practices in 2003.

3.2.2 Clinical quality scores

Quality indicators were developed for the care of diabetes, asthma and angina using a modification of the RAND-UCLA evidence based review criteria (Saxena, 2006; Campbell, 2001; Campbell, 2005; Campbell, 2003) . In each practice random samples of patient medical records over the previous 5 years (or 15

months for some indicators) were extracted for patients over 18 with a diagnosis of diabetes, asthma, or angina. Separate random samples of patient records were drawn in 1998 and 2003 (mean number of records per practice, per condition at each time point between 15 – 21). Clinical quality data obtained in 2001 from a separate study was used for one practice (HES, 2006). Indicators for glycaemic control and use of angiotensin-converting enzyme (ACE) inhibitors along with eye and feet examinations were measured for diabetes. Use of aspirin, beta-blockers and patient exercise capacity were recorded for angina. For asthma patients use of bronchodilator or oral steroids, and peak flow was assessed. Blood pressure, cholesterol and weight were monitored for angina and diabetes, while smoking status and advice was recorded for all conditions. Full details are available elsewhere (Campbell, 2005; Gnanadesigan, 2004). For each condition the quality score for each practice was calculated as the mean percentage of 'necessary' quality indicators met for each patient.

3.2.3 Ambulatory care sensitive admission rates

Data were available on all inpatient (emergency, elective and day case) hospital admissions for angina, diabetes and asthma for each primary care practice from Hospital Episode Statistics (Department of Human Services, 2004) for the financial years 1997/98 – 2004/05. International Classification of Disease (ICD-10) primary diagnostic codes were used to identify the main cause of admission. Healthcare Resource Group (HRG) codes were also used to classify admissions for diabetes (see Tab. **11**). The definitions were based on the Prevention Quality Indicators from the Agency for Healthcare Research and Quality in the United States (AHRQ, 2005), which followed a review by the UCSF-Stanford EPC. The set of quality indicators are consistent with other studies and were the result of a consensus processes involving an expert panel, literature review and empirical methods. The indicators are associated with evidence-based guidelines to reduce complication rates leading to hospital admission (Giuffrida, 1999; Basu, 2002).

Annual practice admission rates were estimated as moving averages to reduce sampling error. For all 60 practices 1998 admission rates were calculated as the total number of admissions over the financial years 1997/98, 1998/99 and 1999/00 divided by the total number of person years at risk of admission in each practice. Similarly, 2003 admission rates were calculated for 43 of the practices using admissions and person years at risk for 2002/03 and 2003/04, and 2004/05.

3.2.4 Size and demographic and socio-economic characteristics of the practice's registered population

For the years 1996 – 2004, data on practice demographic characteristics came from the General Medical Service Statistics database maintained by the Department of Health. A three-year moving average of the practice list size was used to smooth the effects of measurement error in the list size and to align the data better with the financial years used in the Hospital Episode Statistics. The proportion of ethnic minority patients in each practice was attributed to the practice using small area (ward) data from the 1991 and 2001 censuses and the proportion of the practice population resident in different wards.

3.2.5 Measures of health and socio-economic status

The Indices of Multiple Deprivation (IMD) for 2000 and 2004 (GPAQ, 2006) were used to attribute measures of health, education and employment deprivation to the practice population. The health deprivation domain is an area level index of higher than expected numbers of people whose quality of life is impaired by poor health and disability or whose life is cut short by premature death. The education domain is an index of poor educational achievement, and the employment domain identifies those who want to work but are unable to because of unemployment, sickness or disability. Another measure of morbidity in the practice was the age standardised proportion of incapacity benefit and severe disability allowance claimants (Brazier, 2002). The Low Income Scheme Index (LISI) (Office of the Deputy Prime Minister, 2006) for the practice is a measure of low income defined as the percentage of the costs of dispensed prescriptions that are exempt from

prescription charges on the grounds of low income, but not exempt under some other criterion (eg. old age, maternity). This measure is derived from the actual practice population, whereas the IMD domain indices were calculated by attributing ward-based measures to the practice according the proportion of the practice population resident in different wards. From the Allocation of Resources to English Areas (AREA) project (Brazier, 2002) we obtained information about access to secondary care including the average distance from the family practice to the nearest 5 hospitals.

3.2.6 Model specification and variable selection

We estimated separate random-effects longitudinal data models by generalized least squares (GLS) regression for diabetes, asthma and angina admission rates on practice clinical quality scores (see Box 3.1 in the appendix for a formal definition of the model estimated). We risk-adjusted the analysis by controlling for potential confounding factors likely to have a direct effect on hospital admission rates, but be beyond the control of the primary care practice. These included the age and gender distribution of the practice population, the proportion of non-white patients, the IMD health, education, and employment deprivation indices, the proportion of incapacity and severe disability allowance claimants, and the distance to 5 nearest hospitals. We included a binary year variable to allow for a common trend in admissions over time due to changes in the provision of primary care such as the introduction of NHS direct, walk in centres or specialist community treatment centres (LISI, 1995; Sutton, 2002). Health Authority indicator variables were used to control for unobserved environmental and geographical factors that influence admissions across areas. We assumed that structural, institutional and organisational characteristics of the GP practice, for example the number of GPs and nurses, the age and experience of the GPs only influence admission rates indirectly through our measure of clinical quality. We expected the effect of quality on admission rates to be negative.

The random-effects estimator includes a practice specific intercept in the error term to allow for unobserved heterogeneity in the levels of responses across practices, and adjusts the variance for the dependence in admission rates within

practices over time. The model assumes there is no correlation between the practice specific intercept and included covariates. The estimated coefficients account for both the cross-sectional (between practices) and the longitudinal (within practice) associations. *Robust* (Huber/White) and *cluster* standard errors that adjust for heteroskedasticity and relax the assumption of independence in residuals within practices by grouping their residuals were estimated using Stata/SE 9.2. We model the logarithm of admission rates which is equivalent to a Poisson regression model for counts (Powers, 2000).

To test the assumption that quality is uncorrelated with unobserved practice characteristics that have a direct effect on admissions we estimate a Mundlak or within-groups model, (Mundlak, 2004; Skrondal, 2004) by including the mean practice specific quality score over all time periods (between practice effect) as a separate covariate along with the deviation in quality from the practice specific mean quality (within practice effect). The within practice effect is the effect of a change in a practice's quality score over time on the change in practice admission rates and is an unbiased estimate of the effect of quality provided there is no significant change over time in unobserved practice specific characteristics associated with admission rates and quality. A Wald test of the significance of the difference between the two coefficients provides a test for omitted variable bias.[44] To test whether practice attrition was significantly associated with admission rates we performed a variable addition test for attrition bias by including a binary variable indicating whether or not a practice was followed up after 1998 (see Box 3.1 and Box 3.2 in the appendix for a formal approach to this issue). The within practice effect will also be unaffected by attrition bias provided practice drop out was not the result of changes in unobserved factors over time (Verbeek, 1992).

We excluded non-significant covariates at the 10% significance to improve the precision of our estimates by increasing the number of degrees of freedom and reducing collinearity between variables. A backwards stepwise selection procedure was used with F-tests to assess joint significance of variables and to test the validity of our exclusion restrictions. We retained some variables in the model

if they contributed to explaining the variation according to the adjusted R-squared. The RESET test and added variable plots were used to check model functional form. Variance inflation factors were calculated to detect collinearity between variables (Fox, 1997). Weighted estimation by size of practice was used to check sensitivity to imprecisely measured data.

3.3 Results

3.3.1 Effects of GP practice quality on hospital admission rates

Tab. 13 shows mean admission rates and quality scores for each condition in 1998 and 2003. Mean admission rates for diabetes, asthma and angina in 1998 were 14 per 10,000 (95% confidence interval 11.7 to 16.3), 15.6 per 10,000 (12.6 to 18.5), and 18.3 per 10,000 (16.3 to 20.3) respectively. Mean admission rates decreased in 2003, but not significantly, while mean quality scores increased significantly over time for all conditions. The presence of a downward trend in admission rates and an upward trend in quality does not imply that better quality of clinical care is associated with lower admission rates. By including a common time trend variable in the regression model we ensure that any association is not a spurious relationship between two variables trending in the same or opposing directions over time.

Tables 11-13 display the estimated coefficients, 95% confidence intervals and p-values for the multivariate regression analysis relating quality of care to diabetes, asthma, and angina admissions. After controlling for practice characteristics, quality of care for diabetes was associated with increased admissions for diabetes ($p < 0.01$). A 10% increase in the quality score for diabetes care was associated with an 9.5% (2.2%, to 16.8%) increase in diabetes admission rates. For asthma, there was a small and non-significant ($p = 0.78$) effect of quality of care on admission rates (0.5% increase in admissions for a 10% increase in quality; 95% CI, -2.7% to 3.6%. For angina, there was a significant negative effect ($p < 0.002$) of quality on admission rates, with a reduction in admission rate of -6.3% (-10.2% to -2.3%) for each percentage point increase in quality. Taking other

factors into account, the percentage of the variation in log admission rates explained by quality of care was 3.1% for diabetes and 6.1% for angina.

3.3.2 Effects of potential confounding factors on practice admission rates

We controlled for confounding factors to improve the precision of the estimated effects for quality. We find, however, significant and plausible associations between admission rates for the three conditions and demographic, education deprivation, income deprivation, and geographic variables. These factors explained 35%, 64% and 39% of the variation in log admission rates for diabetes, asthma and angina respectively, considerably more of the variation than could be explained by differences in quality of care between practices. The intraclass correlation, which measures the degree of unexplained dependence in admission rates within practices over time or the proportion of unexplained variation due to the practice specific intercept, was 0.06 for both diabetes and asthma, and 0.28 for angina. Hence, the majority of the unexplained variation was within practices over time, suggesting it was due to factors beyond the control of the practice or pure randomness.

3.3.3 Robustness of estimates

We find no evidence of significant omitted variable bias in our random-effect estimates of the effect of quality. The Mundlak or 'within-groups' estimates of the effect of quality were not significantly different from the 'between-groups' estimates. A 10% increase in quality for diabetes within a practice significantly increased ($p = 0.002$) practice admission rates by 15.4% (5.50% to 23.5%). For asthma the within practice effect of quality was non-significant ($p = 0.18$) (a 10% increase in quality within a practice increased admission rates by 3.8%; 95% CI -1.8% to 9.4%). A 10% increase in quality of care for angina within a practice significantly decreased ($p = 0.02$) admission rates by -5.9% (-10.9% to -8.8%). Variable addition tests for attrition bias found no significant difference in admission rates for practices lost to follow up ($n = 17$) at the 5% level for angina,

but for asthma and diabetes admission rates were significantly higher in practices lost to follow up. We found no significant difference in the change in admission rates over time for practices lost to follow up (follow up admission data for n = 14 practices). There was no significant difference, however, in diabetes (difference = 2.94%, p = 0.47) or asthma (difference = -8.5%, p = 0.15) quality scores for practices lost to follow up. The Mundlak or 'within-groups' estimates will be robust to attrition bias provided attrition was not related to changes in unobserved characteristics that affect admission rates over time. Weighted (size of practice) estimation yielded similar results.

3.3.4 Using ACSC admission rates to predict below average quality of care

Given the significant negative association between angina clinical quality and admission rates for angina, we investigated the extent to which angina admission rates can be used to detect below average quality of care in our sample of practices (see Tab. 17). We use indirectly standardized admission rate ratios (actual admission rates divided by expected admission rates from the random-effects model of log angina admission rates on all significant control variables). We defined poorly performing practices to be those where actual admission rates are greater than expected admission rates (indirectly standardized rate ratio > 1), and investigated the sensitivity and specificity of using this criterion to identify observations with below average quality scores (< 67.8%). Fig. 28 shows the angina quality scores for all 102 observations plotted against the indirectly standardized rate ratio. The horizontal line at 67.8% indicates the mean quality score. The vertical line at 1 partitions practices into those with higher than expected admissions (right half) and lower than expected admissions (left half). The proportion of observations with below average quality scores correctly classified by an indirectly standardized rate ratio > 1 (sensitivity) is 63% (see Table 5). The proportion of observations with above average quality scores correctly classified by an indirectly standardized rate ratio of ≤ 1 (specificity) is 57%. The likelihood ratio of a positive test (sensitivity/(1-specificity)) is 1.47.

3.4 Discussion

Our analysis provides mixed evidence that better clinical quality in primary care reduces hospital admission rates for ambulatory care sensitive conditions. Better clinical quality of care significantly reduced hospital admissions for angina by a non-trivial magnitude. However, we found no significant effect of quality on admissions for asthma and a significant positive effect for diabetes, with improved clinical care associated with an increase in the number of admissions.

There are a number of reasons why improved care could lead to increased admissions. In diabetes for example, tighter glycaemic control could result in an increase in the number of admissions for hypoglycaemic complications, or result in admissions for conditions that are detected as a result of good care. Better clinical quality of care may also result in increased case-finding. It is also possible that process measures of quality may reflect unobserved severity of illness amongst diabetes patients: patients with complications are more likely to attend the physician frequently, and therefore may be more likely to have process based quality indicators recorded. Our clinical quality scores were calculated on a selected sample of patients who had been diagnosed with the condition. If the proportion of undiagnosed patients in the practice is related to the practice quality score and the admission rate then our estimates of the effect of quality could be biased. However, our estimates would only be biased by significant changes in unobserved severity of illness or the proportion of undiagnosed patients within practices over time. Our measure of clinical quality may not capture all dimensions of primary care that influence hospital admissions for diabetes.

Our results do not support ACSC admission rates as sensitive or specific measures of clinical quality in primary care. For angina, we did find that quality of care explains a small percent (7.8%) of the total variation in admission rates, but not for asthma and diabetes where the effect was the reverse of that expected. Even for angina, above average ACSC admissions had unimpressive sensitivity (63%) and specificity (57%) as an indicator of below average care. It may be more appropriate to use ACSC admission rates to monitor changes within primary care practices over time in order to remove unobserved confounding factors.

Our findings suggest that considerable caution is required when using ACSC admission rates as a measure of clinical quality in primary care. Morbidity, demographic, socio-economic and geographical factors explain the most variation in admission rates, and a large proportion of the variation was left unexplained. The relationship of ACSC admission rates to quality of primary care is less strong, and may not always even be in the expected direction.

Appendix 2

Box 3.1 Model specification

We specify the following outcome models for practice hospital admission rates:

$$y_{it}^{c} = \beta_0^c + \beta_1^c t + \beta_2^c q_{it}^c + \beta_{3k}^c x_{kit} + \varepsilon_i^c + \varepsilon_{it}^c \qquad [3.2]$$

where $i = 1,\ldots,60$; $t = 1998, 2003$; $c = diabetes, asthma, angina$; y_{it}^c is the admission rate for practice i, in period t, for condition c; t is a time trend dummy variable for the years 1998, and 2003; q_{it}^c the measure of practice quality; and x_{kit} are a set of potential confounding variables that may influence quality and have a direct effect on admission rates, but are considered beyond the influence of the GP practice such as disease prevalence, morbidity in the practice population, age, gender, ethnicity, and socio-economic characteristics of the practice's registered population.

It is assumed that all unobserved time invariant and time varying factors affecting practice admission rates $\left(\varepsilon_i^c \; and \; \varepsilon_{it}^c\right)$ are independent of our measure of quality.

To test the assumption that the effect of quality is not confounded by unobserved, time invariant practice characteristics ε_i^c, we estimate a Mundlak estimator that includes the mean practice quality score $\overline{q}_i^c$ (Mundlak, 1978).

$$y_{it}^{c} = \beta_0^c + \beta_1^c t + \beta_{2\,between}^c \overline{q}_i^c + \beta_{2\,within}^c \left(q_{it}^c - \overline{q}_i^c\right) + \beta_{3k}^c x_{ki} + \varepsilon_i^c + \varepsilon_{it}^c \qquad [3.2]$$

The coefficient $\beta_{2\,within}$ estimates the effect that a change in quality has on a given practice, and will be uncorrelated with ε_i^c. We test whether there is a significant difference between the estimated coefficients β_2^c in [3.1] $\beta_{2\,within}^c$ in [3.2].

Box 3.2 Mundlak regression as a 'test for omitted variable bias'

Mundlak offers a way to model the correlated effects, i.e. the correlation between the time-constant omitted variables (α_i) with the explanatory variables x_i :

$$\alpha_i = \pi \overline{x_i} + w_i$$

In general, if the Mundlak hypothesis is invalid , i.e. $\pi = 0$, this implies that we can use RE estimator.

If the random intercept is thought to reflect the effects of omitted variables, then the assumption that the practice specific random intercept is uncorrelated with quality (variable of interest) will be violated if the omitted variable z_i is correlated with quality:

$$\alpha_i = \gamma_1 z_i + \varphi_i$$

If we model the correlation between z_i and quality as a linear regression we get:

$$z_i = \alpha_0 + \alpha_1 \left(q_{ij} - \overline{q}_i \right) + \alpha_2 \overline{q}_i + u_{ij}$$

Now α_1 will be equal to zero as z_i is assumed time invariant. Hence,

$$z_i = \alpha_0 + \alpha_2 \overline{q}_i + u_i$$

Therefore

$$\alpha_i = \gamma_1 \left(\alpha_0 + \alpha_2 \overline{q}_i + u_i \right) + \varphi_{ij} = \gamma_1 \alpha_0 + \gamma_1 \alpha_2 \overline{q}_i + \gamma_1 u_i + \varphi_i = \delta_0 + \delta_1 \overline{q}_i + \tilde{\varphi}_i$$

Hence we can write the model for admissions y_{it} as

$$y_{it} = \beta_0 + \delta_1 \overline{q}_i + \beta_1 q_{it} + \tilde{\varphi}_i + \varepsilon_{ij} = \beta_0 + \left(\delta_1 + \beta_1 \right) \overline{q}_i + \beta_1 \left(q_{it} - \overline{q}_i \right) + \tilde{\varphi}_i + \varepsilon_{ij}$$

continued

Therefore testing the difference between the coefficients on the between quality effect and the within quality effect is testing whether the bias δ_i is significant. This is why we referring to the Mundlak approach as a test for omitted variables. Some economists tend to think they have to make a choice between a fixed effects regression and a random effects regression. But this shows a failure to understand that one can obtain consistent estimates of the true effect β_1 just by including the 'cluster mean' of that variable (and the cluster means for other variables that appear to be biased). Of course this still does not guarantee the effect is unbiased. If the unobservable covariates that have changed over time are also correlated with the change in quality then the effect will be biased. We try to overcome that problem by including a fixed time trend, but this still relies on the assumption that the change in the omitted variable is common across all practices.

One can also view the test as investigating whether or not it is valid to pool the between and within effects. In some circumstances the between and within effects can have quite different meanings and interpretations. For instance, if we are interested in the effects of age on self assessed health. The between effect will tell us why someone who is x years old reports different self assessed health (SAH) than someone who is a different age. The within effect tells us the effect of aging x number of years on SAH. Conceptually one probably does not want to pool these two estimates. The between effect tells us about cohort or generational effects in reporting health status (older generations may be more stoical) whereas the within effect gives you the effect of getting older on self-assessed health.

To some Mundlak model can be interpreted either as an omitted variable bias test or a test of the 'poolability' of the between and within effects.

By pooling these two kinds of information (by applying RE model) , we obtain more efficient as well as consistent estimates. In our analysis, this is true at least for the variable of interest (i.e. quality).

Tab. 8. Type 2 Diabetes - Criteria used in the clinical audit, (n=1111) (Source: Campbell et al., 2002)

	Frequency met*	95% CI	Sample
Past 14 months, record of:			
Blood pressure	92.9%	91.4 to 94.4	1111
HbA_{1c}	87.1%	85.3 to 89.2	1111
Weight	82.2%	79.9 to 84.5	1111
Serum creatinine	78.8%	76.2 to 81.1	1111
Examination of fundi or visual acuity	71.9%	69.3 to 74.6	1111
Urine proteinuria	69.2%	66.4 to 71.9	1111
Recording of peripheral pulses or record of vibration sense	64.6%	61.6 to 67.2	1111
Visual examination of the feet	62.6%	59.7 to 65.3	1111
Record of hypoglycaemia symptoms if patient on sulphonylurea	*21.9%*	*19.0 to 24.7*	*818*
Past 5 years, record of:			
Smoking status	86.5%	84.4 to 88.4	1111
Documentation of education about diabetes if diagnosed <5 years	*82.9%*	*77.6 to 85.6*	*364*

Serum cholesterol	75.3%	72.8 to 77.8	1111
Advice given to smokers	*62.8%*	*56.2 to 70.4*	*177*
Blood pressure: (criteria developed before publication of UKPDS trial[38])			
Under 80 years: Offered treatment if average of last 3 readings shows diastolic >100 or systolic >150 and diastolic >90	*64.6%*	*55.2 to 74.1*	*99*
Treatment (criteria developed before publication of UKPDS trial[38])			
Referral to a specialist where serum creatinine is >200 mmol/l	*81.2%*	*62.1 to 100*	*16*
For patients under 70, where the last HbA_{1c} was >9, patient offered a therapeutic intervention aimed at improving glycaemic control	*75.9%*	*68.9 to 83.3*	*134*
For patients over 70, where the last HbA_{1c} was >10, patient offered a therapeutic intervention aimed at improving glycaemic control	*64.3%*	*58.8 to 89.6*	*31*
If patient is being treated for hypertension and has proteinuria (macro- but not micro-albuminuria), the patient is on an ACE inhibitor	*50.4%*	*46.2 to 54.6*	*530*
If patient was started on ACE inhibitor, creatinine and potassium were measured within 1 month of starting treatment	*37.5%*	*31.7 to 43.3*	*272*

*Frequency with which care was provided and recorded for patients for whom the necessary care was clinically indicated;
^number of patients to whom the criterion applied. Conditional variables are shown in italics.
Note: these criteria were devised by expert panels using a systematic process to combine evidence with expert opinion.[24]

Tab. 9. Asthma - Criteria used in the clinical audit, (n=1133) (Source: Campbell et al., 2002)

	Frequency met	95% CI	Sample
Past 14 months, record of:			
Record of daily, nocturnal or activity limiting symptoms	40.8%	37.9 to 43.7	1133
Past 5 years, record of:			
Speech rate, pulse rate or respiratory rate during a consultation for an exacerbation of asthma if immediate bronchodilator therapy was used	*100%*	*23.2 to 53.8*	*39*
Oral steroids prescribed if peak flow <60% of normal/predicted	*84.1%*	*75.8 to 93.4*	*65*
Smoking status	80.0%	77.5 to 82.2	1133
Normal or predicted peak flow or record of difficulty using a peak flow meter	74.7%	72.2 to 77.2	1133
Peak flow during a consultation for an exacerbation of asthma	*69.2%*	*64.5 to 73.8*	*376*
Action taken if patient experiencing nocturnal symptoms	*64.8%*	*59.1 to 70.4*	*278*
Smoking advice to smokers	*58.4%*	*52.8 to 64.9*	*253*
Referral to a respiratory physician where oral steroids are used in maintenance treatment	*53.8%*	*34.7 to 73.0*	*26*
Action taken if patient experiencing activity limiting symptoms	*50.6%*	*42.4 to 58.3*	*153*
Inhaler technique	50.1%	47.2 to 53.1	1133
For patients with recorded exercise induced bronchospasm, short acting bronchodilators prescribed for use before exercise	*39.2%*	*32.3 to 46.0*	*194*
Self-management plan for those on high dose steroids or who have had inpatient treatment for asthma	*37.3%*	*30.9 to 43.7*	*217*

Tab. 10. Angina - Criteria used in the clinical audit, (n=1048) (Source: Campbell et al., 2002)

	Frequency met	95% CI	Sample
Past 14 months, record of:			
Blood pressure	85.1%	82.9 to 87.3	1048
Prescribed or advised to take aspirin unless record of contraindication or intolerance	74.0%	71.3 to 76.6	1048
Prescribed ß blocker as maintenance therapy if sole therapy	*51.8%*	*45.6 to 57.9*	*255*
Frequency or pattern of angina attacks	41.5%	38.5 to 44.5	1048
Action taken on blood pressure if systolic blood pressure >160, or >140 if cholesterol level >5.5 mmol/l	*37.1%*	*29.4 to 45.6*	*139*
Exercise capacity	36.3%	33.4 to 39.1	1048
Past 5 years, record of:			
Smoking status	82.3%	80.0 to 84.7	1048
Smoking advice to smokers	*72.3%*	*65.8 to 78.9*	*181*
Cholesterol	62.3%	59.4 to 65.2	1048
Weight advice if overweight	*59.3%*	*54.6 to 64.0*	*425*
Dietary advice	56.9%	53.9 to 59.9	1048
Action taken if cholesterol >5.5 mmol/l	*55.2%*	*47.3 to 62.6*	*162*
Ever recorded:			
Referred for specialist assessment	76.5%	73.9 to 79.1	1048
Referred for an exercise ECG	41.7%	38.7 to 44.7	1048

Tab. 11. Definitions and codes for classifying ambulatory care sensitive condition admissions

Condition	**ICD-10 and HRG codes**[†]	**Hospital Episode Statistics fields**
Diabetes	**Primary ICD-10 diagnoses:** E10 (Insulin-dependent diabetes mellitus); E11 (Noninsulin-dependent diabetes mellitus); E13 (Other specified diabetes mellitus); E14 (Unspecified diabetes mellitus); **or HRG code:** K11 (Diabetes with hypoglycaemic emergency > 69 or with complications); K12 (Diabetes with hypoglycaemic emergency < 70 without complications); K13 (Diabetes with hyperglycaemic emergency > 69 or with complications); K14 (Diabetes with hyperglycaemic emergency < 70 without complications); K15 (Diabetes with other hyperglycaemic disorder > 69 or with complications); K16 (Diabetes with other hyperglycaemic disorder < 70 without complications); K17 (Diabetes with lower limb complications); **or HRG code:** Q15 (Amputations) **with additional (2nd to 7th) ICD-10 diagnoses:** E10; E11; E13; E14	*dismeth* ≠ 8 (Defines a finished consultant spell) *admimeth* ≠ 81 (Excludes hospital transfers) *diag_1* (Primary diagnosis field indicating main cause of admission) diag_2 to diag_7 (additional diagnostic fields) *hrglate* (Healthcare Resource Group)
Asthma	**Primary ICD-10 diagnosis:** J45 (Asthma); J46 (Status asthmaticus)	*dismeth* ≠ 8 (Defines finished consultant spell) *admimeth* ≠ 81 (Excludes hospital transfers) *diag_1* (Primary diagnosis field indicating main cause of admission)
Angina	**Primary ICD-10 diagnosis:** I20 (Angina pectoris); I24 (Other acute ischemic heart diseases) **exclude OPCS-4**[‡] **codes:** K (heart procedures)	*dismeth* ≠ 8 (Finished consultant spell) *admimeth* ≠ 81 (exclude hospital transfers) *diag_1* (Primary diagnosis field indicating main cause of admission) *oper nn* (Operation codes)

[†] International Classification of Disease Codes (ICD-10) and Healthcare Resource Group (HRG) codes. Both fields used for diabetes to identify diabetes related lower limb complications, diabetes related glycaemic complications, and diabetes related amputations

[‡] Operation Procedure Codes Version 4

Tab. 12. Set practice explanatory variables

Control Variables	**Mean**	**SD**	**Range**	**Source**
Practice list female (%)	49.40	2.25	42.66-60.42	GMS statistics
Practice registered patients aged between 0-4 years (%)	13.22	3.36	7.55-24.41	GMS statistics
Practice registered patients aged between 25-44 years (%)	22.81	5.12	7.90-33.52	GMS statistics
Practice registered patients aged between 45-64 years (%)	14.41	3.88	7.69-29.32	GMS statistics
Practice registered patients aged between 65 years (%)	15.45	5.57	3.32-33.40	GMS statistics
Practice population from ethnic minority (%)	11.71	15.33	0.18-63.67	1991/2001 censuses
Index of employment deprivation (Factor score)	12.85	7.64	3.00-50.00	IMD
Index of education deprivation (Factor score)	11.01	19.72	-1.20-87.64	IMD
Average distance from practice to 5 nearest providers (km)	23.62	9.57	13.46-54.33	AREA project
Age standardised incapacity benefit/severe disability allowance (Ratio)	103.09	48.67	41.75-220.94	IMD

Tab. 13. Mean admission rates and quality scores for diabetes, asthma and angina over time

ACSCs	**Mean admission rates per 10,000 person years (95% CI) [SE]**		**Mean quality scores % (95% CI)* [SE]**	
	1998 (n= 60)	**2003 (n = 43)**	**1998 (n= 60)†**	**2003 (n = 43)**
Diabetes	14.22 (11.90 to 16.55) [1.16]	11.52 (9.13 to 13.90) [1.19]	71.09 (67.49 to 74.69) [1.80]	78.66 (75.52 to 81.79) [1.55]
Asthma	15.59 (12.63 to 18.54) [1.48]	12.46 (9.13 to 15.79) [1.64]	58.19 (53.41 to 62.98) [2.39]	70.08 (65.42 to 74.74) [2.31]
Angina	18.31 (16.31 to 20.31) [1.00]	15.97 (13.46 to 18.48) [1.24]	60.09 (57.48 to 62.71) [1.31]	78.33 (75.35 to 81.30) [1.47]

† For angina n = 59 in 1998

Tab. 14. Diabetes – effects of quality on practice admission rates: Generalised Least Squares regression estimates (Random Effects)*

Variable	Coefficient (95% CI) (Percentage effects)**	*P*- Value
Primary care quality score (%)	0.95 (0.02 to 1.68)	0.01
Time trend‡		
2003	-36.80 (-58.90 to -14.71)	0.001
Demographic characteristics§		
Practice population from ethnic minority (%)	0.21 (0.08 to 0.34)	0.001
Practice registered patients aged between 0-4 years (%)	-1.27 (-1.87 to -0.68)	<0.001
Health Deprivation~		
Age standardised incapacity benefit/severe disability allowance (Ratio)	-1.28 (0.73 to 1.84)	<0.001
Socio-economic characteristics‖		
Index of employment deprivation (Factor score)	-0.20 (-0.43 to 0.04)	0.10
Access to secondary care¶		
Average distance from practice to 5 nearest providers (km)	13.70 (6.34 to 21.07)	<0.001
Average distance from practice to 5 nearest providers squared (km^2)	-0.19 (-0.31 to -0.07)	0.10
Health Authority effects††		
South Essex	48.45 (4.44 to 92.47)	0.03
West Pennine	41.0 (1.02 to 80.98)	0.04
Bury and Rochdale	36.81 (-8.80 to 82.41)	0.11
Avon	71.9 (40.8 to 103.1)	<0.001
	Adjusted $R^2 = 0.45$	
	Intra-cluster correlation = 0.06	

*Dependent variable diabetes admission rate per 10,000 person years (logs) for years 1998 and 2003. Multivariate regression analysis was based on n=103.

**Percentage effect calculated as $(d\log y/dx)$/mean of x for continuous variables and $(d\log y/dx)*100$ for indicator variables.

†Coefficient is the percentage change in practice admission rates due to a 1% increase in the average practice quality score (measured in %).

‡ Coefficient measures proportionate difference in practice admission rates between 2003 and 1998.

§Coefficient measures the percentage change in practice admission rates due to a 1% increase in 1. the average population from ethnic minority (measured as %); 2. the average population aged 0-4 (measured in %) compared with population aged 5 and over (measured in %)

~ Coefficient measures the percentage change in practice admission rates due to a 1% increase in the age standardised rate ratio of patients receiving incapacity benefit or severe disability allowance.

‖Coefficient measures the percentage change in practice admission rates due to 1% increase in the average percentage of the costs of dispensed prescriptions that are exempt from prescription charges on the grounds of low income, *ceteris paribus*.

¶Coefficients measure nonlinear relationship (n-shaped) between practice admission rates and the average distance from GP practice to the 5 nearest secondary care providers (measured in kilometers), *ceteris paribus*. Point of inflexion at 50 km. Elasticity around the mean distance is 1.24.

††Coefficients measures proportionate difference in admission rates for practices in each health authority compared with practices in Somerset, and Enfield and Haringey Health Authorities. South Essex Health Authority is in South East England; West Pennine and Bury and Rochdale Health Authorities are in North West England; Avon Health Authority is in South West England.

Tab. 15. Asthma – effects of quality on practice admission rates: Generalised Least Squares regression estimates (Random Effects)*

Variable	Coefficient (95% CI) (Percentage effects)**	P- Value
Dimension of quality †		
Family practice quality score (%)	0.05 (-0.27 to 0.36)	0.78
Time trend‡		
2003	-25.28 (-43.03 to -7.53)	0.005
Practice demographic characteristics§		
Practice list female (%)	2.50 (1.03 to 3.97)	0.001
Practice registered patients between 25-44 years (%)	-0.66 (-1.46 to 0.14)	0.11
Practice registered patients between 45-64 years(%)	0.47 (-0.05 to 0.97)	0.08
Socio-economic characteristics¦¦		
Index of employment deprivation (Factor score)	0.29 (0.13 to 0.46)	<0.001
Access to secondary care¶		
Average distance from practice to 5 nearest providers (Km)	9.17 (3.36 to 14.99)	0.002
Average distance from practice to 5 nearest providers squared ($Km^{2)}$	-0.13 (-0.23 to -0.04)	0.004
Health Authority effects††		
West Pennine	65.71 (0.54 to 0.98)	<0.001
Bury and Rochdale	59.42 (34.46 to 84.39)	<0.001
	Adjusted R^2 =0.63	
	Intra-cluster correlation = 0.06	

*Dependent variable diabetes admission rate per 10,000 person years (logs) for years 1998 and 2003. Multivariate regression analysis was based on n=103.

**Percentage effect calculated as (dlog$y$/dx)/mean of x for continuous variables and (dlog$y$/dx)*100 for indicator variables.

†Coefficient measures percentage change in practice admission rates due to 1% increase in the average practice quality score (measured in %)

‡ Coefficient measures proportionate difference in practice admission rates between 2003 and 1998.

§Coefficient measures percentage change in practice admission rates due to 1% increase in 1. the average practice list female (measured in %); 2. the average percentage of population in respective age groups compared with percentage aged 0-24 and 65 and over.

¦¦Coefficient measures percentage change in practice admission rates due to 1% increase in the average percentage of the costs of dispensed prescriptions that are exempt from prescription charges on the grounds of low income*s*.

¶Coefficients measure nonlinear relationship (n-shaped) between practice admission rates and the average distance from GP practice to the 5 nearest secondary care providers (measured in kilometers). Point of inflexion at 35 km. Elasticity around the mean distance is 0.45.

**Coefficients measures proportionate difference in admission rates for practices in each health authority compared with practices in South Essex, Somerset, Enfield and Haringey, and Avon Health Authorities. West Pennine and Bury and Rochdale Health Authorities are in North West England.

Tab. 16. Angina – effects of quality on practice admission rates: Generalised Least Squares regression estimates (Random Effects)*

Variable	Coefficient (95% CI) (Percentage effects)**	*P*- Value
Dimension of quality†		
Family practice quality score (%)	-0.63 (-1.02 to -0.23)	0.002
Demographic characteristics‡		
Practice population from ethnic minority (%)	-0.17 (-0.30 to -0.04)	0.01
Practice list female (%)	-1.65 (-3.53 to 0.24)	0.09
Practice list patients aged 65 years and over (%)	0.44 (0.17 to 0.72)	0.001
Socio-economic characteristics§		
Index of employment deprivation (Factor score)	0.19 (-0.016 to 0.39)	0.03
Index of education deprivation	0.04 (-0.014 to 0.09)	0.15
Access to secondary care‖		
Average distance from practice to 5 nearest providers (Km)	0.09 (0.02 to 0.15)	0.01
Average distance from practice to 5 nearest providers squared ($Km^{2)}$	-0.002 (-0.003 to -0.001)	0.002
Health Authority effects¶		
South Essex	-32.05 (-57.87 to -6.22)	0.01
Bury and Rochdale	-25.17 (-51.18 to 0.85)	0.07
Avon	-19.44 (-0.44 to 0.051)	0.09
	Adjusted R^2 = 0.38	
	Intra-cluster correlation = 0.28	

*Dependent variable angina admission rate per 10,000 person years (logs) for years 1998 and 2003. Multivariate regression analysis was based on n=102 due to missing data for a practice quality score.

**Percentage effect calculated as (dlog$y$/dx)/mean of x for continuous variables and (dlog$y$/dx)*100 for indicator variables.

† Coefficient measures the percentage change in practice admission rates due to a 1% increase in the average practice quality score (measured in %)

‡ Coefficient measures the percentage change in practice admission rates due to a 1% increase in 1. the average population from ethnic minority (measured as %); 2. the average practice list female (measured in %); 3. the average population aged 65 (measured in %) compared with population aged 0-64 (measured in %).

§Coefficient measures the percentage change in practice admission rates due to a 1% increase in 1. the average education score; 2. the average percentage of the costs of dispensed prescriptions that are exempt from prescription charges on the grounds of low income .

‖Coefficients measure nonlinear relationship (n-shaped) between practice admission rates and the average distance from GP practice to the 5 nearest secondary care providers (measured in kilometers). Point of inflexion at 22.5 km. Elasticity around the mean distance is -0.12.

¶Coefficients measures proportionate difference in admission rates for practices in each health authority compared with practices in West Pennine, Somerset, and Enfield and Haringey Health Authorities.

Tab. 17. Sensitivity and specificity of indirectly standardized admission rate ratios from model for angina admissions as a predictor of below average quality of angina care.

	Number of practices with lower than expected admission rate	**Number of practices with higher than expected admission rate quality**
Number of practices with above average quality	True negative (TN) = 31	False negative (FP) = 23
Number of practices with below average quality	False positive (FN) = 18	True positive (TP) = 30
	Sensitivity = 63%[*]	**Specificity = 57%**[†]

* Sensitivity is (TP/(TP+FN)) x100. The proportion of practices with quality scores below the mean value (score<68) with an indirectly standardized admission rate ratio > 1.

[†]Specificity **is** (TN/(FP+TN)) x100. The proportion of practices with quality scores above the mean value (score>68) with an indirectly standardized rate ratio < or equal to 1.

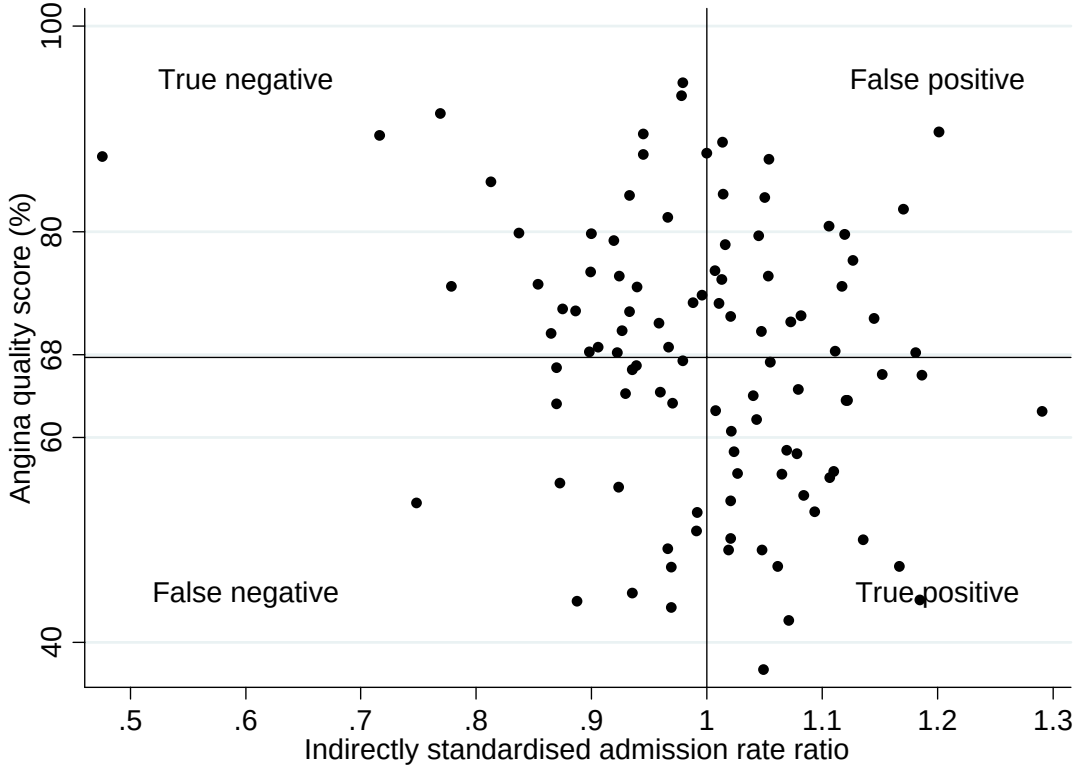


Fig. 28legend. Angina quality scores (percent) against indirectly standardized admission rate ratios (actual/expected) (n = 102). Horizontal line indicates mean practice quality score (68%). Observations to the right of vertical line have higher than expected admission rates.

Do File for estimates - STATA commands, version 9SE

```
//DIABETES
/*Random effects estimator*/

xi: xtreg logascratex10000diabampma diabscore i.year employmentscore
r_ibsda popprop04 ethnicmin health_au_2 health_au_3 health_au_4
health_au_6 acutdistn5 acutdistn5sq, re robust cluster(pract_id)

mfx compute, dyex

/*Display adjusted R-squared*/
display 1 - (1-e(r2_o))*((e(N)-1)/(e(N) - e(df_m) + 1))

xttest0

xi: xtreg logascratex10000diabpma avgdiabscore deltadiabscore i.year
popprop04 ethnicmin r_ibsda health_au_2 health_au_3 health_au_4
health_au_6 acutdistn5 acutdistn5sq, re robust cluster(practices)
mfx compute, dyex at(deltadiabscore = 74.2493)

test avgdiabscore =  deltadiabscore
lincom  deltadiabscore - avgdiabscore

xi: xtreg logascratex10000diabpma avgdiabscore deltadiabscore i.year
avgpopprop04 deltapopprop04 avgethnicmin deltaethnicmin avgLISI
deltaLISI health_au_2 health_au_3 health_au_4 health_au_6 acutdistn5
acutdistn5sq, re robust cluster(practices)
mfx compute, dyex at(deltadiabscore = 74.2493)


lincom  avgdiabscore - deltadiabscore
lincom avgpopprop04 - deltapopprop04
lincom avgethnicmin - deltaethnicmin
lincom avgLISI - deltaLISI


/*Testing for nonresponse bias*/
xi: xtreg logascratex10000diabpma diabscore i.year popprop04 ethnicmin
r_ibsda health_au_5 health_au_3 health_au_4 health_au_6 acutdistn5
acutdistn5sq dropout, re robust cluster(pract_id)
mfx compute, dyex

xi: xtreg diabscore i.year dropout, re robust cluster(pract_id)
mfx compute, eydx

/*testing whether practices who drop out have a different trend in
admission rates*/

/*Testing for nonresponse bias*/
xi: xtreg logascratex10000diabpma i.year*dropout, fe robust
cluster(pract_id)
/*Testing for nonresponse bias*/
xi: xtreg logascratex10000diabpma i.year, fe robust cluster(pract_id)

/*Estimates combining non-response and omitted variable bias*/
xi: xtreg logascratex10000diabpma avgdiabscore deltadiabscore i.year
popprop04 ethnicmin r_ibsda health_au_5 health_au_3 health_au_4
health_au_6 acutdistn5 acutdistn5sq dropout, re robust cluster(pract_id)

mfx compute, dyex at(deltadiabscore = 74.2493)

test avgdiabscore =  deltadiabscore
lincom  deltadiabscore – avgdiabscore
```

```
//ASTHMA

xi: xtreg logascratex10000asthma asthscore employmentscore i.year
popprop_F popprop2544 popprop4564 health_au_3 health_au_4 acutdistn5
acutdistn5sq, re robust cluster(practices)

mfx compute, dyex

/*Display adjusted R-squared*/
display 1-  (1-e(r2_o))*((e(N)-1)/(e(N) - e(df_m) + 1))

xttest0

/*Testing for omitted variable bias*/

xi: xtreg logascratex10000asthma avgasthscore deltaasthscore i.year
popprop_F popprop2544 popprop4564 employmentscore health_au_3 health_au_4
acutdistn5 acutdistn5sq, re robust cluster(practices)
mfx compute, dyex at(deltaasthscore = 63.1533)

test avgasthscore  =  deltaasthscore

lincom  deltaasthscore - avgasthscore

/*Display adjusted R-squared*/
display 1-  (1-e(r2_o))*((e(N)-1)/(e(N) - e(df_m) + 1))

xi: xtreg logascratex10000asthma avgasthscore deltaasthscore i.year
deltaeducationscore avgeducationscore avgpopprop_F deltapopprop_F
avgpopprop2544 deltapopprop2544 avgpopprop4564 deltapopprop4564
health_au_3 health_au_4 acutdistn5 acutdistn5sq, re robust
cluster(practices)

lincom  avgasthscore - deltaasthscore
lincom deltaeducationscore - avgeducationscore
lincom avgpopprop_F - deltapopprop_F
lincom avgpopprop2544 - deltapopprop2544
lincom avgpopprop4564 - deltapopprop4564


/*Testing for non-response bias*/

xi: xtreg logascratex10000asthma asthscore i.year popprop_F popprop04
popprop2544 employmentscore  health_au_3 health_au_4 acutdistn5
acutdistn5sq dropout, re robust cluster(practices)
mfx compute, dyex

xi: xtreg logascratex10000asthma avgasthscore deltaasthscore i.year
popprop_F popprop04 popprop2544 employmentscore  health_au_3 health_au_4
acutdistn5 acutdistn5sq dropout, re robust cluster(practices)
mfx compute, dyex at(deltaasthscore = 63.1533)

test avgasthscore  =  deltaasthscore

lincom  deltaasthscore - avgasthscore

xi: xtreg asthscore i.year popprop_F popprop04 popprop2544
employmentscore  health_au_3 health_au_4 acutdistn5 acutdistn5sq dropout,
re robust cluster(practices)
mfx compute, dyex

/*Testing for nonresponse bias*/
xi: xtreg logascratex10000asthma i.year*dropout, fe robust
cluster(pract_id)
```

```
//ANGINA

/*Use random effects estimator*/

iis pract_id

xi: xtreg logascratex10000angma angscore i.year employmentscore popprop_F
popprop65 ethnicmin educationscore acutdistn5 acutdistn5sq health_au_2
health_au_4 health_au_6,  re robust cluster(practices)

mfx compute, dyex

/*Display adjusted R-squared*/
display 1-  (1-e(r2_o))*((e(N)-1)/(e(N) - e(df_m) + 1))

xttest0

/*weighted estimates*/
iis pract_id

xi:  xtreg  logascratex10000angma  angscore  employmentscore  popprop_F
popprop65 ethnicmin educationscore acutdistn5 acutdistn5sq health_au_2
health_au_4 health_au_6, re robust cluster(practices)

mfx compute, dyex

/*Display adjusted R-squared*/
display 1-  (1-e(r2_o))*((e(N)-1)/(e(N) - e(df_m) + 1))
xttest0

/*Note LISI included because improves fit of the model - adjusted R-
square*/

/*Testing for omitted variable bias*/

xi: xtreg logascratex10000angma avgangscore deltaangscore employmentscore
popprop_F   popprop65 ethnicmin educationscore acutdistn5 acutdistn5sq
health_au_2 health_au_4 health_au_6, re robust cluster(practices)
mfx compute, dyex at(deltaangscore = 67.7811 )

test avgangscore =  deltaangscore
lincom  deltaangscore - avgangscore

xi:  xtreg  logascratex10000angma  avgangscore  deltaangscore  avgLISI
deltaLISI     deltaeducationscore     avgeducationscore     avgethnicmin
deltaethnicmin  avgpopprop_F  deltapopprop_F  avgpopprop65  deltapopprop65
health_au_3    health_au_4    acutdistn5    acutdistn5sq,    re    robust
cluster(practices)

lincom  avgangscore -  deltaangscore
lincom deltaeducationscore - avgeducationscore
lincom deltaethnicmin - avgethnicmin
lincom avgpopprop_F - deltapopprop_F
lincom avgpopprop65 - deltapopprop65
lincom avgLISI - deltaLISI


/*Testing for nonresponse bias*/
iis pract_id

xi:  xtreg   logascratex10000angma  angscore  LISI  popprop_F   popprop65
ethnicmin educationscore acutdistn5 acutdistn5sq health_au_2 health_au_4
health_au_6 dropout,  re robust cluster(practices)
mfx compute, dyex
```

```
/*Testing for nonresponse bias*/
xi:    xtreg    logascratex10000angma    i.year*dropout,    fe    robust
cluster(pract_id)

/*Generating predicted values for angina*/

xi:  xtreg    logascratex10000angma  employmentscore  popprop_F  popprop65
ethnicmin educationscore acutdistn5 acutdistn5sq health_au_2 health_au_4
health_au_6,  re robust cluster(practices)

capture drop expamdxb
capture drop expamdxbu
capture drop residu

predict expamdxb if angscore!=., xb
predict expamdxbu if angscore!=., xbu
predict residu, u

capture drop stdamdratesxb
capture drop stdamdratesxbu

gen stdamdratesxb = logascratex10000angma/ expamdxb
gen stdamdratesxbu = logascratex10000angma/ expamdxbu

//Scatterplot ANGINA
twoway  (scatter  angscore  stdamdratesxbu  if  angscore  ~=  .,  sort
msymbol(circle)                msize(small)                mcolor(black)
mlabcolor(black)yline(67.78107,                        lstyle(foreground))
xline(1,lstyle(foreground))), /*
*/  ytitle(Angina  quality  score  (%),  size(medsmall)  color(black))
xtitle(Indirectly  standardised  admission  rate  ratio  ,  size(medsmall)
color(black))   graphregion(fcolor(white)   lcolor(white)   lwidth(vthin)
ifcolor(white) /*
*/ ilcolor(white) ilwidth(vthin)) plotregion(fcolor(white) lcolor(white)
lwidth(vthin) ifcolor(white) ilcolor(white) ilwidth(vthin)) text(95 0.6
"True  negative",  place(c))  text(45  0.6  "False  negative",  place(c))
text(45  1.2  "True  positive",  place(c))  text(95  1.2  "False  positive",
place(c)) /*
*/ xlabel(0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3) ylabel(40 60 68 80 100)

//Sensitivity and Specificity
//TP
list practice if angscore < 67.78107  &  stdamdratesxb >1 & angscore!=.
//FN
list  practice  if  angscore<  67.78107  &   StandardisedAdmissionRates<1  &
angscore!=.
//TN
list  practice  if  angscore>  67.78107  &   StandardisedAdmissionRates<1  &
angscore!=.
//FP
list  practice  if  angscore>  67.78107  &   StandardisedAdmissionRates>1  &
angscore!=.

capture drop qualclass admclassxb admclassxbu

gen qualclass = 0 if angscore > 67.78107
replace qualclass = 1 if angscore < 67.78107
gen admclassxb = 1 if stdamdratesxb > 1
replace admclassxb = 0 if stdamdratesxb < 1
gen admclassxbu = 1 if stdamdratesxbu > 1
replace admclassxbu = 0 if stdamdratesxbu < 1
gen admclassru = 0 if residu > 0
replace admclassru = 1 if residu < 0
```

CONCLUSION

A large body of the literature has shown that physicians can, and likely do, influence resource allocation 1) by supplying differentiated nonretradable services, or 2) by choosing the level of a non-contractible input (quality or effort) or, finally, 3) by influencing the patient's demand through persuasion. Whenever the physician influence patient' demand, standard theory of the competitive firm can not be used to analyze the behaviour of the practitioner. The consumer sovereignty is no longer the normative justification for resource allocation and the willingness to pay cannot be based on the consumers' exogenous preferences, but on physician preferences at the individual and collective level.

Institutional environments, market conditions, regulatory settings and, the last but not least, intrinsic motivations and social norms are important drivers of doctors behaviour. Doctors behaviour may be determined by all these aspects and modified by incentives they carry on. However, the implications of incentive approach on physician practice style and ethical codes is not clear since it will depend on how the tension between external incentives and interior values is solved. A strong system of ethics may attenuate, or totally remove, the incentives to provide an opportunistic behaviour. Conversely, external incentives may positively ('crowd in') or negatively ('crowd out') impact on doctor's moral motivation. Then, the contribution of external incentives to quality improvement will only be maximised if we understand their impact on the internal drivers of doctors. The new UK GP contract provides a case in point. On the one hand, the incentivisation (agreed with the profession itself) of indicators about the routine treatment of single chronic conditions –such as diabete, asthma and angina - does make sense in the context of the evidence. On the other hand, it seems possible that, as an increasing proportion of total doctor work is incentivised, the risks of crowding out of motivation to perform the non-incentivised more complex or simply caring tasks is increased. This is an area of policy that really does need to be supported by high quality evidence. The bulk of research has been on the impact of payment systems on behaviour of physician in post and has ignored the implications for the medical labour market.

Financial incentives definitely influence performance both on the negative and positive side. A great number of studies from health economics literature dealt with this topic. Yet, many empirical studies are based on natural experiments and are therefore opportunistic. Only a few studies have been able to control accurately for the many other differences between GPs that may influence their behaviour. The most notable omissions are patient characteristics, case mix and GP characteristics. These omissions are particular relevant from a theoretical perspective. There appears to be general agreement in the literature on the key differences between fee-for-service, capitation and salary in terms of their key incentives and that mixed payment systems present several advantages over pure retrospective and prospective payment systems. Quality incentives are generally implicit in this systems. To alter this equation, quality-related performance pay has been introduced in a number of health plans. In this case, the regulator establishes an explicit link between quality and the physician's income. The major problem recognised by the general literature is that a 'pure' payment for performance system leads to difficulties in presence of multitasking. In turns, the problem of multitasking further reinforces the argument in favour of mixed payment systems. Specifically, mixed payment reduces distortions in effort allocation when pay-for-performance is imperfect by stimulating noncontractible dimensions of quality that complement treatment. Additionally, accurate measure for physician actions that promote quality are extremely difficult to quantify. Using appropriate measures for quality is essential in designing a pay-for-performance system. More generally, the problem of finding good measures for quality impairs empirical works. The evidence contained in chapter III tried to contribute to fill in this gap by assessing the validity of 'hospitalisations for ambulatory care sensitive conditions' as a measure of clinical (technical quality).

Our analysis provides mixed evidence that better clinical quality in primary care reduces hospital admission rates for ambulatory care sensitive conditions (ACSCs). Better clinical quality of care significantly reduced hospital admissions for angina by a non-trivial magnitude. However, we found no significant effect of quality on admissions for asthma and a significant positive effect for diabetes,

with improved clinical care associated with an increase in the number of admissions.

Our results do not support ACSC admission rates as sensitive or specific measures of clinical quality in primary care. For angina, we did find that quality of care explains a small percent (6.1%) of the total variation in admission rates, but not for asthma and diabetes where the effect was the reverse of that expected. Even for angina, above average ACSC admissions had unimpressive sensitivity (63%) and specificity (47%) as an indicator of below average care. It may be more appropriate to use ACSC admission rates to monitor changes within primary care practices over time in order to remove unobserved confounding factors.

Our findings suggest that considerable caution is required when using ACSC admission rates as a measure of clinical quality in primary care. Morbidity, demographic, socio-economic and geographical factors explain the most variation in admission rates, and a large proportion of the variation was left unexplained. The relationship of ACSC admission rates to quality of primary care is less strong, and may not always even be in the expected direction.

References

Agency for Healthcare Quality and Research (AHQR), *AHRQ Quality Indicators—Guide to Prevention Quality Indicators: Hospital Admission for Ambulatory Care Sensitive Conditions*. Rockville, MD. Pub. No. 02-R0203. 2004. Available at: http://www.qualitytools.ahrq.gov/qualityreport. Accessed March 21, 2006.

Alchain A. A. and Demsetz H., (1972) 'Production, Information Costs and Economic Organization', *American Economic Review*, 62, 777-795.

Amundson, G., L. I. Solberg, M. Reed, E. M. Martini, and R. Carlson. (2003),. 'Paying for quality improvement: Compliance with tobacco cessation guidelines'. *Joint Commission Journal on Quality and Safety* 29 (2): 59-65.

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

Aveyard P. (1997), 'Assessing the performance of general practices caring for patients with asthma'. *Br J Gen Pract*; 47(420): 423-426.

Barigozzi F., and Levaggi R. (2005), 'New Developments in Physician Agency: the Role of Patient Information' Working Papers 550, Dipartimento Scienze Economiche, Università di Bologna.

Basu J, Friedman B, Burstin H. 'Primary Care', HMO Enrollment and Hospitalization for ACS Conditions: A New Approach', *Med Care*, 40 (12); 1260-1269.

Bauchner H., Simpson L. and Chessare J. (2001), 'Changing physician behaviour'. *Arch. Dis. Child*.; 84(6): 459-462.

Baum, A., Newman, S., Weinman, J., West, R. and mcManus, C. (1997); *Cambridge Handbook of* Psychology, Health, and Medicine, Cambridge: Cambridge University Press.

Begley C. E., Slater C. H., Engel M. J. and Reynolds T. F. (1994), 'Avoidable hospitalizations and socio-economic status in Galveston County, Texas'. *J Community Health*; 19(5): 377-387.

Bickerdyke, I., Dolamore, R., Monday, I. and Preston, R. 2002, 'Supplier-Induced Demand for Medical Services', Productivity Commission Staff Working Paper, Canberra November.

Billings J. and Hasselblad V. (1989), 'A preliminary study: Use of small area analysis to assess the performance of the outpatient delivery system in New York City'. *Health Systems Agency of New York City*.

Billings J., Anderson G. M. and Newman L. S. (1996), 'Recent findings on preventable hospitalizations'. *Health Aff (Millwood)*; 15(3): 239-249.

Billings J., Zeitel L., Lukomnik J., Carey T. S., Blank A. E. and Newman L. (1993), 'Impact of socioeconomic status on hospital use in New York City'. *Health Aff (Millwood)*; 12(1): 162-173.

Bindman A. B., Grumbach K., Osmond D., Komaromy M., Vranizan K., Lurie N., Billings J. and Stewart A. (1995), 'Preventable hospitalizations and access to health care'. *JAMA*; 274(4): 305-311.

Blustein J, Hanson K, Shea S. Preventable hospitalizations and socioeconomic status. *Health Aff* 1998; 17:177-89.

Boerma, W., Dubois, C. A., Rico, A. and Saltman, R. (2004), Mapping PC organization in Europe. In R. Saltman, A. Rico and W. Boerma (eds).

Bourgueil Y., Marek A., Mousquès J. (2005), '*La participation des infirmières aux soins primaires dans six pays européens en Ontario et au Québec',* IRDES. Rapport.

Brazier J, Roberts J, Devrill M.(2002) 'The estimation of a preference-based measure of health from the SF-36'. *J Health Econ.* 2002;21:271-292.

Brown L. J. and Barnett J. R. (1992), 'Influence of bed supply and health care organization on regional and local patterns of diabetes related hospitalization'. *Soc Sci Med*.; 35(9): 1157-1170.

Bunker J. and Brown B. (1974), 'The Phisician patient as an Informed Consumer of Surgical Services', *New England Journal of Medicine*, 1051-1055.

Campbell S, Martin O Roland, Elizabeth Middleton and David Reeves, (2005) 'Improvements in quality of clinical care in English general practice 1998-2003: longitudinal observational study, British Medical Journal';331: 1121-1126

Campbell S, Roland M, Middleton E, Reeves D. (2005) 'Improvements in the quality of clinical care in English general practice 1998-2003: longitudinal observational study'. *BMJ.*; 331: 1121-1125.

Campbell S, Steiner A, Robison J, Webb D, Raven A, Roland M. (2003) 'Is the quality of care in general medical practice improving? Results of a longitudinal observational study'. *Br J Gen Pract.*;53:298-304.

Campbell S. M., Braspenning J., Hutchinson A. and Marshall M. (2002), 'Research methods used in developing and applying quality indicators in primary care'. *Qual Saf Health Care*; 11(4): 358-364.

Campbell S. M., Hann M., Hacker J., Burns C., Oliver D., Thapar A., Mead N., Gelb Safran D. and Roland M. (2001), 'Identifying predictors of high quality care in English general practice: observational study'. *Br Med J* ; 323 7316 784 -787

Campbell S. M., Roland M. and Buetow S. (2000), 'Defining quality of care'. *Soc Sci Med* ; 51(11): 1611 -1625.

Campbell SM, Hann M, Hacker J, Roland MO.(2002) 'Quality assessment for three common condition in primary care: validity and reliability of review criteria developed by expert panels for angina, asthma and type 2 diabetes'. *Qual Saf Health Care*.;11:125-130.

Campbell SM, Roland MO, Shekelle PG, Cantrill JA, Buetow SA, Cragg DK, 'Development of review criteria for assessing the quality of management of stable angina, adult asthma and non insulin dependent diabetes mellitus in general practice'. *Qual Health Care*.; 8:6-15.

Caplin A. and Leahy J. (2001), 'Psychological Expected Utility Theory and Anticipated Feelings'. *Quarterly Journal of Economics*; 116(1): 55-80.

Casanova C. and Starfield B. (1995), 'Hospitalizations of children and access to primary care: a cross-national comparison'. *Int J Health Serv*.; 25(2): 283-294.

Casanova C., Colomer C. and Starfield B. (1996), 'Pediatric hospitalization due to ambulatory care-sensitive conditions in Valencia (Spain)'. *Int J Qual Health Care*; 8(1): 51-59.

Chaix-Couturier C., Durand-Zaleski I., Jolly D. and Durieux P. (2000), 'Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues'. *Int J Qual Health Care*; 12: 133-142.

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

Chalkley M. and Malcomson J. M. (2000), 'Government Purchasing of Health Services', in Culyer A. J. and Newhouse J. P. (eds.); Vol 1A: 847-890.

Che Y. and Yoo S. (2001), 'Optimal Incentives for Teams'. *The American Economic Review*; 525-541.

Cromwell, J., Mitchell, J., (1986) 'Physician-induced demand for surgery'. *Journal of Health Economics* 5, 293–313.

Culler S. D., Parchman M. L. and Przybylski M. (1998), 'Factors related to potentially preventable hospitalizations among the elderly'. *Med Care*; 36(6): 804-817.

Culyer A. and Newhouse J. P. (2000), *The Handbook of Health Economics,* Amsterdam, North Holland.

Culyer A. J. (1989), "The normative economics of health care finance and provision", Oxford Review of Economic Policy 5:34-58.

Dafny , Leemore , Gruber and Jonathan (2005), 'Public insurance and child hospitalizations: access and efficiency effects'. *Journal of Public Economics*; 89(1): 109-129.

Davidson S., Stephen, M. Manheim, S. Werner, S, Hohlen M, Yudkowsky (1992), 'Prepayment with office-based physicians in publicly direction' in Anthony J. Culyer and Joseph P. Newhouse, 1: 1175-1200.

De Jaegher K. and Jegers M. (2001), 'The Physician-Patient relationship as a game of strategic information transmission'. *Health Economics*; 10: 651-668.

Deci E., Koestner R. and Ryan R. (1999), 'A meta-analytical review of experiments examining the effects of extrinsic rewards on intrinsic motivation'. *Psychol Bull*; 125: 627–668.

Department of Human Services (2004), *The Victorian Ambulatory Care Sensitive Conditions Study*, *2001–02*, Public Health, Rural and Regional Health and Aged Care Services Division, Victorian Government Department of Human Services Melbourne Victoria.

Department of Human Services. *The Victorian Ambulatory Care Sensitive Conditions Study, 2001–02*. Public Health, Rural and Regional Health and Aged Care Services Division, Victorian Government

Dixit A. (2002), 'Incentives and Organizations in the Public Sector: An Interpretive Review'. *Journal of Human Resources*.

Donabedian A. (1980), *The Definition of Quality and Approaches to its Assessment.* Ann Arbor, MI: Health Administration Press.

Donaldson C. (1999), 'Valuing the benefits of publicly-provided health care: Does 'ability to pay' preclude the use of 'willingness to pay'?' *Social Science and Medicine*; 49:551-563.

Donaldson C., Gerard K. (1989), Countering moral hazard in public and private health care systems: A review of recent evidence. Journal of Social Policy 1989; 18:235-251.

Donaldson C., Gerard K. (1993), *Economics of Health Care Financing* Macmillian, Basingstoke.

Donatini A., Rico A., D'Ambrosio M., *et al* (2001), *Health Care Systems in Transition: Italy,* Rico A, Cetani T (eds). Copenhagen: European Observatory on Health Care Systems. Available at www.euro.who.int/document/e73096.pdf

Doran T., Fullwood C., Gravelle H., Reeves D., Kontopantelis E., Hiroeh U. and Roland M. (2006), 'Pay-for-Performance Programs in Family Practices in the United Kingdom', *N. Engl J Med*; 355(4): 375-384.

Dranove D. (1988), 'Demand inducement and the physician/patient relationship'. *Economic Inquiry*; 26: 251–298.

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

Dranove D. and Satterthwaite M. (2000), 'The Industrial Organization of Health Care Markets', in Culyer A. J. and Newhouse J. P. (eds.), Vol. 1B: 1093-1139.

Dranove D. and White W. (1994), 'Recent Theory and Evidence on Competition in Hospital Markets'. *Journal of Economics and Management Strategy*; 3(1): 169–209.

Dranove D. and White W. D. (1987), 'Agency and the Organization of Health Care Delivery'. *Inquiry*; 24: 405-415.

Durojaiye L. I., Hutchison T. and Madeley R. J. (1989), 'Improved primary care does not prevent the admission of children to hospital'. *Public Health*; 103(3): 181-188.

Dusheiko M. and Gravelle H. (2005), draft 'The effect of Personal Medical Service contracts on admission rates', draft National Primary Care Research and Development Centre CHE, University of York.

Dusheiko M. Gravelle H., Jacobs R. and Smith P. (2003), 'The Effects of Budgets on Doctor Behaviour: Evidence from a Natural Experiment', University of York, DERS Discussion Paper 03104.

Eggleston K. (2005), 'Multitasking and mixed systems for provider payment'. *Journal of Health Economics*; 24(1): 211-223.

Eisemberg J.M. (1986), *Doctors' Decision and the Cost of Medical Care*, Health Administration Press, Ann Arbor, MI.

Ellis, R. P. and McGuire, T. G., (1986),. 'Provider behavior under prospective reimbursement : Cost sharing and supply', *Journal of Health Economics,* 5(2): 129-151.

Ellis R. P. and McGuire T. G. (1990), 'Optimal Payment Systems for Health Services'. *Journal of Health Economics*; 9: 375-396.

Ellis R. P. and McGuire T. G. (1993), 'Supply-side and demand-side cost sharing in health care', Journal of Economic Perspectives 7:135-151.

Ellis, R. P. (1998), 'Creaming, skimping and dumping: provider competition on the intensive and extensive margins1', *Journal of Health Economics*, 17(5): 537-555

Encinosa III W., Gaynor M. and Rebitzer J. B. (2002), 'The Sociology of Groups and the Economics of Incentives: Theory and Evidence on Compensation Systems'. Agency for Health Care Research and Quality, *Working paper.*

Epstein A. M. (2006), 'Paying for Performance in the United States and Abroad'. *N Engl J Med*; 355: 406-408.

Evans R. G. (1974), 'Supplier-Induced Demand: Some Empirical evidence and Implication', in Perlman M. (ed): 162–173.

Fairbrother G., Hanson K. L., Friedman S. and Butts G. C. (1999), 'The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates'. *American Journal of Public Health*; 89(2): 171-175.

Farley P. J. (1986), 'Theories of the price and quantity of physician services'. *Journal of Health Economics*; 5: 315–333.

Fattore G. and Torbika A. (2004), 'Health service benefit catalogues in Europe. Country report. Italy'. Accessed October 28, 2006. http://www.ehma.org/_fileupload/File/Projects/Benefit_Report_Italy.pdf.

Feldman R., Sloan F., 1988. Competition among physicians, revisited. In: Greenberg, W. (Ed.), Competition in the Health Care Sector: Ten Years Later. Duke University Press, London, p.252.

Feldstein M. S. (1970), 'The Rising Price of Physicians' Service'. *Review of Economics and Statistics*; 51: 121-133.

Ferguson B., Gravelle H., Dusheiko M., Johns R. and Sutton M. (2002), 'Benchmarking for performance monitoring in the new NHS: variations in practice admission rates'. *Journal of Health Services Research and Policy*; 7: 170-176.

Fireman, B, Bartlett, J, Selby, J. Can Disease Management Reduce Health Care Costs By Improving Quality? Health Affairs 2004; 23: 63 – 75.
Folland, Goodman, Stano. (2001), *The Economics of Health and Health Care*, 3rd ed. Prentice Hall.

Fox J. (1997), *Applied Regression Analysis, Linear Models, and Related Methods*, Thousand Oaks, CA: Sage.

France G., Taroni F. and Donatini A. (2005), 'The Italian health-care system'. *Health Economics*; 14 (Suppl1): 187-202.

Frey B (1997), *Not Just for the Money; an Economic Theory of Personal Motivation* Cheltenham, UK and Brookfield, US: Edward Elgar Publishing.

Fuchs V.R.(1978), "The supply of surgeons and demand for operations", The Journal of Human Resources 35-56.

Gaynor M. (1994), 'Issues in the industrial organization of the market for physician services'. *The Journal of Economics and Management Strategy*; 39(1): 211–255.

Gaynor M. (2004), 'Competition and quality in health care markets: what do we know, what don't we know?'. *Paper* commissioned by the Federal Trade Commission, mimeo, Department of Public Policy, Carnegie Mellon University, Pittsburgh.

Gaynor M. and Gertler P. (1995), 'Moral Hazard and Risk Spreading in Partnerships'. *The RAND Journal of Economics;* 26(4): 591-613.

General Practice Assessment Questionnaire. Available at: http://www.gpaq.info/. Accessed March 21, 2006.

Giuffrida A. and Gravelle H. (2001), 'Inducing or restraining demand: the market for night visits in primary care'. *Journal of Health Economics*; 20: 755-799.

Giuffrida A.**,** Gravelle H. and Roland M. (1999), 'Measuring quality of care with routine data: avoiding confusion between performance indicators and health outcomes'. *BMJ*; 319: 94-98.

Gnanadesigan N, Saliba D, Roth D, Solomon DH, Chang JT, Schnelle J. et al. The quality of care provided to vulnerable older community-based patients with urinary incontinence, *J Am Med Dir Assoc* 2004; 5: 141-146.

Goddard M., Davies H. T. O., Dawson D., Mannion R. and McInnes F. (2002a), 'Clinical performance measurement: part 1-getting the best out of it'. *Journal of the Royal Society of Medicine*; 95: 508–510.

Goddard M., Davies H. T. O., Dawson D., Mannion R. and McInnes F. (2002b), 'Clinical performance measurement: part 2-avoiding the pitfalls'. *Journal of the Royal Society of Medicine*; 95: 508–510.

Gosden T., Forland F. , Kristiansen I. S., Sutton M., Leese B., Giuffrida A., Sergison M. and Pedersen L. (2001), 'Impact of payment method on behaviour of primary care physicians: a systematic review'. *Journal of Health Services Research & Policy*; 6(1): 44–55.

Gosden T., Pedersen L. and Torgerson D. (1999), 'How should we pay doctors? A systematic review of salary payments and their effect on doctor behaviour'. *Q J Med*; 92: 47–55.

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

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

Griffiths C., Naish J., Sturdy P. and Pereira F. (1996), 'Prescribing and hospital admissions for asthma in east London'. *BMJ*; 312(7029): 481-482.

Harrison S. and Smith C. (2004), 'Trust and moral motivation: redundant resources in health and social care?'. *Policy and Politics*; 2: 371–386.

Hay J. and Leahy M.J. (1982), "Physician induced-demand", Journal of Health Economics 2:231-244.

Hayward R. A., Bernard A. M., Freeman H. E. and Corey C. R. (1991), 'Regular source of ambulatory care and access to health services'. *Am J Public Health*; 81(4): 434-438.

Hearnshaw H. M., Harker R. M., Cheater F. M., Baker R. H. and Grimshaw G. M. (2003), 'Are audits wasting resources by measuring the wrong things? A survey of methods used to select audit review criteria'. *Qual Saf Health Care*; 12: 24-28.

HFA (2005) http://www.istat.it/sanita/Health/. Accessed Accessed March 21, 2006

Hickson B., Altemeier A. and Perrin M. (1987), 'Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study'. *Pediatrics*; 80: 344–350.

Hillman A.L., Welch W. and Pauly M. (1992), "Physicians' utilization and charges for outpatient diagnostic imaging in a Medicare population", Journal of the American Medical Association 268:2050-2054.

Hillman et al. (1990), "Frequency and costs of diagnostic imaging in office practice – a comparison of self-referring and radiologist-referring physician", New England Journal of Medicine 323:1604-1608.

Holmstrom B. e P. Milgrom (1991), Multitask Principal-Agent Analyses:

Hospital Episode Statistics. Available at: http://www.hesonline.nhs.uk/. Accessed March 21, 2006.

Ittner C. D., Larcker D. F. and Pizzini M. J. (2004), 'Performance-based pay in professional service firms'. *Working Paper,* University of Pennsylvania.

Ifo's Database for Institutional Comparisons in Europe (DICE), **http://www.cesifogroup.de/portal/page?_pageid=36,34773&_dad=portal&_schema=PORTAL**; Accessed March 21, 2006

Jackson G. and Tobias M. (2001), 'Potentially avoidable hospitalisations in New Zealand, 1989-98'. *Aust N Z J Public Health*; 25(3): 212-219.

Jankowski R. (1999), 'What do hospital admission rates say about primary care?'. *BMJ*; 319: 67-68.

Johnston G. and Lynn R. (2004), 'Did the invisible hand rock the cradle? An investigation of children's hospitalisations in New Zealand'. *Journal of Health Services Research and Policy*; 9, suppl. 2 (4): 23-28.

Kessler D. and McClellan M. (1996), "Do doctors practice defensive medicine?" Quarterly Journal of Economics 111:353-390.

Koszegi B. (2003), 'Health, Anxiety and Patient Behavior'. *Journal of Health Economics*; 22(6): 1073-1084.

Kouides R., Bennet N., Lewis B., Cappuccio J., Barker W. and LaForce M. (1998), 'Performance based physician reimbursement and influenza immunisation rates in the elderly'. *American Journal of Preventive Medicine*; 14: 89–95.

Kozsegi B. (2005), 'Emotional Agency'. *Quarterly Journal of Economics*, 22 , 1073-1084

Krakauer H., Jacoby I., Millman M. and Lukomnik J. E. (1996), 'Physician impact on hospital admission and on mortality rates in the Medicare population'. *Health Serv Res*; 31(2): 191-211.

Krasnik A. and Gottschau A. (1993), 'Doctor and patient characteristics as modifiers of the effect of a changing remuneration system in general practice'. *Danish Medical Bulletin*; 40: 380–382.

Krasnik A., Groenewegen P., Pedersen P., von Scholten P., Mooney G.., Gottschau A. et al. (1990), 'Changing remuneration systems: effects on activity in general practice'. *BMJ*; 300: 1698–1701.

Kristiansen I. and Hjortdahl P. (1992), 'The general practitioner and laboratory utilisation: Why does it vary?'. *Family Practice;* 9: 22–27.

Kristiansen I. and Holtedahl K. (1993), 'The effect of the remuneration system on the general practitioner's choice between surgery consultations and home visits'. *J Epidemiol Community Health;* 47: 481–484.

Kristiansen I. and Mooney G. (1993), 'The general practitioner's use of time: is it influenced by the remuneration system?'. *Social Sci Med*; 37: 393–399.

Kuhn M. (2003), *Quality in Primary Care: Economic approaches to analysing quality-related physician behaviour*, Office of Health Economics.

Labelle, R. G. Stoddart and T. Rice (1994), 'A Re-examination of the Meaning and Importance of Supplier-induced Demand',*Journal of Health Economics;* 13(3): 347-68.

Laditka J. N. (2004), 'Physician Supply, Physician Diversity, and Outcomes of Primary Health Care for Older Persons in the United States'. *Health and Place*; 10: 231-244.

Laditka J. N., Laditka S. B. and Probst J. C. (2005), **'**More May Be Better: Evidence of a Negative Relationship between Physician Supply and

Hospitalization for Ambulatory Care Sensitive Conditions'. *Health Services Research*; 40(4): 1148-1166.

Laditka JN. Hazards of Hospitalization for Ambulatory Care Sensitive Conditions among Older Women: Evidence of Greater Risks for African Americans and Hispanics, *Med Care Res Rev.* 2003;60:468-495.

Lambrew JM, DeFriese GH, Carey TS, Ricketts TC, Biddle AK. The effects of having a regular doctor on access to primary care, *Med Care.* 1996;34:138-51.

Lien H. M., Ma C. T. A. and McGuire T. G. (2004), 'Provider–client interactions and quantity of health care use'. *Journal of Health Economics;* 23: 1261–1283.

Low income scheme index: a new deprivation scale based on prescribing in general practice. *BMJ.* 1995;310:165-169.

Luck J., Peabody J., Dresselhaus T., Martin L. and Glassman P. (2000), 'How Well Does Chart Abstraction Measure Quality? A Prospective Comparison of Standardized Patients with the Medical Record?'. *American Journal of Medicine*; 108(8): 642-649.

Ma C. A. (1994), 'Health Care Payment Systems: Cost and Quality Incentives'. J*ournal of Economics & Management Strategy*; 3(1): 93-112.

Ma C. A. and McGuire T. G. (1997); 'Optimal Health Insurance and Provider Payment'. *American Economic Review;* 87(4): 685-704.

Marrée J, Groenewegen PP. *Back to Bismarck: Eastern European health care systems in transition*. Aldershot: Avebury, 1997.

Marshall M, Campbell S, Hacker J, Roland M. *Quality indicators for general practice. A practical guide to clinical quality indicators for primary care health professionals and managers*. London, Royal Society of Medicine Press; 2002; 15 -21.

Marshall M. and Harrison S. (2005), 'It's about more than money: financial incentives and internal motivation'. *Qual. Saf. Health Care*; 14(1): 4-5.

McElduff P., Lyratzopoulos G., Edwards R., Heller R. F., Shekelle P. and Roland M. (2004), 'Will changes in primary care improve health outcomes? Modelling the impact of financial incentives introduced to improve quality of care in the UK'. *Qual. Saf. Health Care*; 13(3): 191-197.

McGuire T. G. (2000), 'Physician Agency', in Culyer and Newhouse (eds.*)*; Vol.1A: 461-536.

McGuire T.G., Pauly M. (1991), 'Physician response to fee changes with multiple payers'. *J. Health Econ*. 10 4 , 385–410.

McPherson K. Strong PM, Epstein A et al. Regional variation in the use of common surgical procedures within and between England and Wales, Canada and the United States of America. *Social Science & Medicine*, 1981; 15A: p.273-288

Millman M, editor. Committee on Monitoring Access to Personal Health Care Services. Washington, DC: National Academy Press; 1993.

Mitchell J.M., Sass T.R. (2005), Physician ownership of ancillary services: Indirect demand inducement or quality assurance? pp. 263-289(27).

Mooney G. and Ryan M. (1993), 'Agency in health care: getting beyond first principles'. *Journal of Health Economics*; 12: 125-135.

Moore AT and Roland MO. How much of the variation in general practitioner referral rates is due to chance? *BMJ*, 1989; 298:500-502.

Morgenstern H. (1982), 'Uses of ecological analysis in epidemiologic research'. *Am J Public Health*; 72: 1336-1344.

Mundlak Y. (1978), On the pooling of time series and cross section data. *Econometrica*.;46:69-85.

Newhouse P. (1996), 'Reimbursing Health Plans and Health Providers: Efficiency in Production versus Selection'. *Journal of Economic Literature* 34 (3): 1236-1263.

Newhouse, J. P. 2002. *Pricing the priceless:Ahealth care conundrum. TheWalras-Pareto lectures*. Cambridge, MA: MIT Press.

Nguyen N .X. and Derrick F.W. (1997), "Physician behavioural response to a Medicare price reduction", Health Services Research 32:283-298.

NHS Executive, Quality and performance in the NHS: High Level Performance indicators, 1999.

Niti M and Ng TP. Avoidable hospitalization rates in Singapore, 1991-1998: assessing trends and inequities of quality in primary care. *J Epidemiol. Community Health.* 2003; 57:17-22.

OECD. *The reform of health care: A comparative analysis of seven OECD countries*. Paris: OECD, 1992

Office of the Deputy Prime Minister. Indices of Deprivation. http://www.communities.gov.uk/index.asp?id=1128440.

Pappas G., Hadden W. C., Kozak L. J. and Fisher G. F. (1997), 'Potentially avoidable hospitalizations: inequalities in rates between US socioeconomic groups'. *American Journal of Public Health*; 87(5): 811-816.

Parchman M. L. and Culler S. (1994), 'Primary care physicians and avoidable hospitalizations'. *J Fam Pract*; 39(2): 123-128.

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

Pauly M. V. (1979), "The ethics and economics of kickbacks and fee-splitting", Bell Journal of Economics 10:344-352.

Pauly M. V. (1996), 'Economics of Multispecialty Group Practice'. *Journal of Ambulatory Care Management*; 19(3): 26-33.

Pauly M. V. (2000), 'Insurance reimbursement', in: A. J. Culyer & J. P. Newhouse (ed.); vol. 1A: 537-560.

Perlman M. (1974) *The Economics of Health and Medical Care*, Macmillan, London

Petersen L. A., Woodard L. D., Urech T., Daw C. and Sookanan S. (2006), 'Does Pay-for-Performance Improve the Quality of Health Care?'. *Ann Intern Med.*; 145: 265-272.

Phelps C. E. (1992), 'Diffusion of information in medical care'. *Journal of Economic Perspectives*; 6(3): 23-42.

Phelps C.E. (1986), " Induced demand: can we ever Know its extent?" Journal of Health Economics 5:355-365.

Pizzini M. (2005), 'Team-based compensation in professional service firms'. School of Management, The University of Texas at Dallas, Draft: May, 2005.

Powell A. E., Davies H. T. O. and Thomson R. G. (2003), **'**Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls'**,.** *Qual Saf Health Care*; 12: 122-128.

Powers DA, Xie Y. *Statistical Methods for Categorical Data Analysis*. Academic Press. San Diego. 2000.

Rashidian A., Black N. and Russell I. (2005), 'Financial incentives and quality improvement'. *Qual. Saf. Health Care*; 14(3): 227-227.

Reid F. D. A., Cook D. G. and Majeed A. (1999), 'Explaining variation in hospital admission rates between general practices: cross sectional study**'**. *BMJ*; 319(7202): 98–103.

Reinhardt U.(1978), “Comment on monopolistic elements in the market for physician services”, in: Greenberg, ed., Competition in the Health Care Sector (Aspen Publications) 121-148.

Rice T. (1983),. The impact of changing Medicare reimbursement rates on physician-induced demand. Med. Care 21 8 , 803–815.

Rice T.(1998), The Economics of Health Reconsidered (Health Administration Press, Chicago, IL).

Ricketts T. C., Randolph R., Howard H. A., Pathman D. and Carey T. (2001), ‘Hospitalization rates as indicators of access to primary care’. *Health Place*; 7(1): 27-38.

Ritchie L., Bisset A., Russell D., Leslie V. and Thomson I. (1992), ‘Primary and pre-school immunisation in Grampian: progress and the 1990 contract’. *BMJ*; 304: 816–819.

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

Rochaix L. (1993), ‘Financial incentives for physicians: the Québec experience’. *Health Economics*; 2: 163-176.

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.ournal of Health Economics, 8: 53-84.

Roland M. (2004), ‘Linking Physicians' Pay to the Quality of Care - A Major Experiment in the United Kingdom’. *N Engl J Med*; 351(14): 1448-1454.

Rosenthal M. B. and Frank R. G. (2006), ‘What Is the Empirical Basis for Paying for Quality in Health Care?’. *Medical Care Research and Review*; 63 (2): 135-157.

Rosenthal M. B., Frank R. G., Li Z. and Epstein A. M. (2005), 'Early Experience With Pay-for-Performance From Concept to Practice', *JAMA,* 294: 1788-1793.

Rosenthal, M. B., R. Fernandopulle, H. R. Song, and B. Landon (2004), Paying for quality:Providers' incentives for quality improvement. *Health Affairs (Millwood)* 23 (2): 127-41.

Roski, J., R. Jeddeloh, L. An, H. Lando, P. Hannan, C. Hall, and S. H. Zhu. 2003. The impactof financial incentives and a patient registry on preventive care quality: Increasing provider adherence to evidence-based smoking cessation practice guidelines. *Preventive Medicine* 36 (3): 291-99.

Rutstein D. D., Berenberg W., Chalmers T. C., Child C. G., Fishman A. P. and Perrin E. B. (1970), 'Measuring the quality of medical care. A clinical method'. *The New England Journal of Medicine;* 294(11): 582-588.

Saltman R. B., Rico A. and Boerma W. (2004), 'Primary care in the driver's seat?Organizational reform in European primary care'. European Observatory on Health Systems and Policies Series; available at http://www.blackwell-synergy.com/links/doi/10.1111/1467-9515.00360/pdf. Accessed: 11/28/2006.

Saxena S, George J, Barber J, Fitzpatrick J, Majeed, (2006) 'An Association of population and practice factors with potentially avoidable admission rates for chronic diseases in London: cross sectional analysis', *J R Soc Med*;99:81–89.

Schreiber S. and Zielinski T. (1997), 'The Meaning of Ambulatory Care Sensitive Admissions: Urban and Rural Perspectives'. *The Journal of Rural Health Fall*; 13(4): 276-284.

Scott A. and Shiell A. (1997), 'Analyzing the effect of competition on general practitioners' behavior using a multi level modeling framework", *Health Economics*, 6:577-.588.

Skrondal A, Rabe-Hesketh (2004), S.*Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, Fl: Chapman & Hall/CRC.

Smith P. C. and York N. (2004), **'**Quality Incentives: The Case Of UK General Practitioners'. *Health Affairs*; 23 (3): 112-118.

Stano M. (1987), 'A clarification of theories and evidence on supplier-induced demand for physicians' services', *Journal of Human Resources*, 22: 611-620.

Starfield B. (1994), 'Primary Care: Is It Essential?'. *Lancet*; 344: 1129-1133.

Sutton M, Gravelle H, Morris S, Leyland A, Windmeijer F, Dubbing C. (2002), *Allocation of Resources to English Areas: Individual and Small Area Determinants of Morbidity and Use of Health Care*, Report for Department of Health. Available at: http://www.isdscotland.org/isd/info3.jsp?pContentID=2859&p_applic=CCC&p_service=Content.show& . Accessed March 21, 2006.

U.S. Department of Health and Human Services: Agency for Healthcare Research and Quality (AHRQ), The National Healthcare Quality Report (NHQR), 2005.

Verbeek M, Nijman TE.(1992), 'Testing for selectivity bias in panel data models' *International Economic Review*.; 33:681-703

Weinberger M et al.(19956), 'Does increased access to primary care reduce hospital readmissions?', *New England Journal of Medicine*; 334: 1441-7.

Weissman J. S., Gatsonis C. and Epstein A. M. (1992), 'Rates of avoidable hospitalization by insurance status in Massachusetts and Maryland'. *JAMA;* 268(17): 2388–2394.

Wienke G. W., Boerma W. and Dubois C. A., Mapping primary care across Europe, Chapter 2, 22-49 in Saltman.

Yip W.(1998), 'Physician responses to medical fee reductions: changes in the volume and intensity of supply of Coronary, Artery Bypass Graft (CABG) surgeries in the medicare and private sectors', *Journal of Health Economics*, 17: 675-700.

Zweifel P. and Breyer F. (1997), *Health economics*; New York - Oxford; Oxford University Press.

Zweifel, P. and Manning W.G. (2000), Moral Hazard and Consumer Incentives in Health Care, in: Culyer, A.J. und J.P. Newhouse (Eds), 409-459.