Journal of Health Economics 11 (1993) 239-258. North-Holland

Should physicians be permitted to 'balance bill' patients?*

Jacob Glazer

Department of Economics, Boston University, Boston, MA, USA and Faculty of Management, Tel Aviv University, Tel Aviv, Israel

Thomas G. McGuire

Department of Economics, Boston University, Boston, MA, USA

Received March 1992, final version received December 1992

This paper studies the efficiency effects of physician fees when the insurer (possibly the government) pays a fee for each procedure, and the doctor may supplement the fee by an extra charge to the patient, a practice known as 'balance billing.' Monopolistically competitive physicians can discriminate among patients on the basis of both price and quality. Equilibria with and without balance billing are compared. The paper analyzes and recommends a new fee policy, a form of payer 'fee discrimination.'

1. Introduction

Setting physician fees is one of the most pressing issues in U.S. health policy. Led by the federal Medicare program (accounting for about 28 percent of payments to physicians and hospitals (ProPAC, 1989)), and in response to continued rapid increase in health care costs, virtually all payers are resetting fees. Changes in technology and in the relative supply of physicians in different specialties have caused fees to grossly exceed costs for some procedures, and to clearly underpay for others (Cromwell et al., 1989; Hsiao et al., 1988). In the single procedure (among more than 5,000) that accounts for six percent of all physician payments in Medicare, cataract removal and lens implant, the physician was paid by the government an average of more than $1,600 in 1986 for 53 minutes work. See Table 1. In

Correspondence to: Thomas G. McGuire, Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, USA.

*Research for this paper was supported by contract 278-0024 and grant 1K05MH00832-01 from the National Institute for Mental Health. In addition, Glazer received research support from the Tel-Hashomer Hospital. We are grateful to Randy Ellis, Eric Lattimer, Michael Riordan, Bob Rosenthal, and Steve Zuckerman for helpful comments on an earlier draft.

Table 1

Charges and times for twenty most costly procedures Medicare, 1986

Rank	Code	Description	Allowed avg charge	Percent of all charges	Mean time	Std Dev. time
1	66984	Cataract removal, insert lens	$1610.86	5.67	52.7	19.0
2	90060	Office visit/intermed.	23.55	4.43	17.4	4.5
3	90050	Office visit/limited	18.92	3.84	13.8	5.7
4	90260	Hospital visit/interm.	28.41	3.70	17.6	6.3
5	90250	Hospital visit/limited	23.50	2.72	11.0	5.0
6	90220	Hospital care/new, compreh.	70.94	2.01	51.7	22.7
7	90620	Comprehensive consultation	84.40	1.94	74.2	48.2
8	93000	ECG, with report	31.24	1.49	7.6	5.9
9	71020	X-ray exam. of chest	20.45	1.44	3.4	2.1
10	52601	Prostatectomy	1076.45	1.29	62.2	16.8
11	90070	Office visit/extended	30.39	1.16	26.2	9.4
12	90270	Hospital visit/extended	35.73	1.07	NA	NA
13	90040	Office visit/brief	15.71	1.07	34.2	30.4
14	66983	Cataract removal, insert lens	1607.72	1.03	NA	NA
15	90240	Hospital visit/brief	19.36	0.88	9.0	4.7
16	93010	ECG, report only	11.74	0.86	NA	NA
17	90080	Office visit/comprehensive	44.76	0.85	37.8	15.2
18	90020	Office visit/new, compreh.	52.09	0.74	32.9	16.8
19	71010	X-ray exam. of chest	12.47	0.68	NA	NA
20	27130	Total hip joint replacement	2364.03	0.65	127.9	39.0

Notes: List of top twenty Medicare procedures and charge information was provided to us by Nancyanne Causino of the Harvard School of Public Health. Mean time information is from Table 31 of Hsiao et al., (1988). Table 31 also reports the standard error of the mean in terms of percentage of the geometric mean. This was converted to an approximate standard deviation in our table by multiplying the standard error in percentage terms, times the mean, times the square root of the sample size used for Table 31, which was approximately 100 in each specialty surveyed by Hsiao et al. Time estimates by procedure were drawn from the following specialties: Internal Medicine (90060, 90050, 90260, 90220, 93000, 90070, 90080), General Surgery (90250, 90020), Ophthalmology (66984), Psychiatry (90620), Radiology (71020), Urology (52601), Pediatrics (90040), Thoracic Surgery (90240), Orthopedic Surgery (27130).

another high dollar volume surgical procedure, prostatectomy, physician time was 20 percent greater, but the payment averaged only $1,076. If the same physician spent an hour seeing patients in a private office or a hospital, the income would have in most cases been less than $100. Such apparent anomalies, along with the belief among some that fees are simply 'too high' (Pauly, 1991), moved the federal government to support a fee restructuring.[1]

The fees listed in Table 1 are only what Medicare pays to physicians. Subject to some limitations, physicians can also, in a practice known as 'balance billing,' charge an additional price to the patient over and above the government fee. In this case the doctor's revenue is the government fee plus

[1]See the Notice of Proposed Rule Making in th *Federal Register* of June 5, 1991 for a recent description of the reforms.

the price paid by the patient. Some health payers prohibit this practice, forcing the doctor to accept the third-party payer's fee as full payment[2].

Thus, any payer faces two key decisions when setting a fee for doctors' services. The first is the level of the fee, and the second is the yes/no decision about whether to force the provider to accept the fee as full payment, to, in other words, prohibit 'balance billing' of the patient.[3] Fee policy has obvious distributional effects. Higher fees and balance billing transfers some surplus from patients to physicians, an effect emphasized in a recent report by the GAO (1989) evaluating the experience in the New England states. This paper emphasizes the efficiency effects of fee levels and balance billing when the monopolistically competitive physician can discriminate among patients on the basis of both price and quality. We compare the market equilibrium with balance billing to one without, where the doctor must sell at the payer-set fee or not at all.

We show that there is always a fee policy with balance billing that dominates in efficiency terms any fee policy without balance billing. Our analysis of physician price discrimination through balance billing leads us to consider a new form of fee policy for a payer, which we refer to as 'payer fee discrimination.' We show that the payer can reduce the fee paid when the physician charges a price, increase the fee for the fee-only patients, and improve the efficiency of the physician market, at no net cost.

In the health economics literature, the model commonly used to address fee issues portrays a representative physician facing a downward-sloping demand for services (due to product differentiation), with patients' willingness to pay augmented by a payer (usually Medicare) set fee. The model allows for both demand or supply-constrained equilibria (Mitchell and Cromwell, 1982; Zuckerman and Holahan, 1991). The profit-maximizing physician charges a price to high-demand patients, and may serve others up to the point where either patients' marginal benefit falls to zero (and they are unwilling to accept more services even if the government pays the fee) or the physician's marginal cost rises to the fee level.[4] One important flaw in this

[2]Most Blue Shield plans are in this category, as well as state-operated Medicaid programs for the poor. Insurers may require some copayment by the patient, but the physician total payment can still be limited by the payer-set fee. In Medicare, physicians charge a balance bill for about 30 percent of procedures (PPRC, 1990).

[3]Our analysis applies most immediately to the Medicare program, and we will call attention to specific Medicare payment policies at various points in the paper. In Medicare, a physician can sign away the right to balance bill patients in exchange for a premium over the usual fee (and other minor forms of favorable treatment). Historically, a minority of physicians have taken this option. As limits on balance billing discussed later undermine the economic advantage to retaining the privilege, the number of physicians who argee to forego balance billing altogether will increase.

[4]Other papers have been concerned with the overall level of fees. Baumol (1988), reasoning on the basis of a competitive model, makes a long-run argument that an aggressive policy by Medicare limiting fees will restrict supply of services to the elderly. Recognizing that fees are in effect an indemnity insurance payment to the patient as well as a price to providers, Pauly

model is that competition from other physicians plays no role in affecting demand of the 'representative' physician. When the fee is raised, for example, the physician may wish to accept more patients. But other physicians would presumably behave similarly, and this should affect the position of demand for the physician under study. Fee policy is market-wide, not physician-specific. Another problem with the model is that the physician has no choice about 'quality,' effort, or any other variable which may be influenced by fee policy and which may affect patients' valuation.[5] The quality choice is an important one, potentially serving to equilibrate supply with demand when prices are set by regulation, a process explored in research on the fee-regulated airline industry and in other contexts (White, 1981)[6].

The paper is organized as follows. In Section 2 the model is presented and the market equilibrium is described. In Section 3 we analyze the welfare effects of prohibiting balance billing, and in Section 4 we introduce the idea of payer fee discrimination. Section 5 concludes the paper.

2. The model

2.1. Patients: willingness to pay and quality

Two physicians offer a differentiated product. Differentiation is introduced by assuming that patients are distributed uniformly on a line of unit length between the two physicians and demand one unit of service per period, distance serving as a geographic metaphor for patient tastes.[7] The uniformity assumption simplifies the model at no real cost. All of our qualitative results would go through under a more general distribution.

Quality of the service is indicated by a variable s chosen by the physician,

(1991) comes to an opposing conclusion. He argues that Medicare fees are sufficiently high in general that most utilization is not constrained by supply but by demand, and that a decrease in fees will mainly shift rents and not decrease utilization of services. McGuire and Pauly (1991) analyze supply of a single physician, and focus on the potential 'income effect' of a Medicare fee change. Large income effects generated by Medicare fee reductions may lead to an increase in supply to all payers. McGuire and Pauly show that the model of physician 'target income' emerges as income effects become very strong.

[5]Feldman and Sloan (1988) and Wedig et al. (1989) have each considered the effect of fees on the quality of physician services, showing that in a model of a single physician with market power, increasing fees raises quality. These papers do not include competition or price or quality discrimination. Pauly (1991) recognizes that doctors' quality choice can be used to bring costs into line with Medicare fees.

[6]Most of the literature on price and quality discrimination from the broader field of industrial organization is confined to the monopoly case. See, for example, Maskin and Riley (1984) or Tirole (1988). Recently, Katz (1984), Borenstein (1985), Lederer and Hurter (1986) and Holmes (1989) have analyzed models of price discrimination under imperfect competition.

[7]Product differentiation can be introduced in a variety of ways. Since our model includes other complexities, such as quality choice and price and quality discrimination, we have chosen to analyze a simple form of product differentiation and demand. Distance is one reason why patients would differentiate among physicians. Information about particular physicians or other characteristics of their practices may matter as well.

with higher s indicating lower quality. Variable s can be regarded as the degree of physician 'skimping' on quality, and will be discussed in more detail shortly. With these two considerations, demand is represented simply in the following way. A patient at distance t $(0 \leq t \leq 1)$ from a physician offering services at quality s values a unit of service from that physician at $\bar{U} - t - s$.[8]

Quality in this context should be understood to be any costly non-price attribute of health care that affects patients' valuation, including dimensions of convenience, comfort, communication about medical conditions, and other factors, as well as a narrowly defined 'clinical' quality of care (Wedig et al., 1989). In concrete terms, it is easiest to think of quality as the time a physician spends with a patient to conduct a procedure. Regarding the physician's time as an input into production of health and associated other 'outputs' valued by patients, more time leads to a higher quality but more costly encounter. As Tirole (1988) points out, 'quality' and 'quantity' are generally equivalent at a formal level in models of cost and demand. Does a patient enjoy a higher quality procedure when the physician spends more time, or does he get a higher quantity of physician services? Because there is no fundamental distinction between 'better' and 'more,' we stick with the quality interpretation, with physician time being the real-world counterpart to the variable we have in mind.[9]

2.2. *Physicians: price and quality discrimination*

The payer sets a fee f for physician services that applies to both physicians.[10] The physician is assumed to know patients' willingness to pay. Each physician may set a price p over the fee to charge to some patients. For these patients, the physician receives the fee paid by the payer plus the price charged the patient. Other patients are served at the fee. The physician's patients will be divided into two groups, the price-paying patients and the

[8]With only one physician, the patient's reservation price $\bar{U}$ will affect willingness to pay, but with two or more physicians in sufficiently close competition, it is the patient's alternative source of service, not the reservation price, that matters. We therefore disregard other factors, such as income or health status, that may affect $\bar{U}$.

[9]There is good evidence that physicians have discretion about the amount of time they spend with patients. Table 1 reports the mean and (approximate) standard deviations of time estimates for performing the procedures most costly for Medicare. The time estimates come from Hsiao et al. (1988), where investigators provided physicians with brief patient vignettes for each studied procedure. Standard deviations were about half the mean in most cases, indicating clearly that for conducting these procedures on carefully standardized cases, substantial discretion exists in terms of time spent.

[10]Medicare requires beneficiaries to pay a $75 annual deductible and a copayment of 20 percent of the fee. About 75 percent of the beneficiaries have these out-of-pocket costs met by Medicaid or private insurance which supplements the Medicare coverage. We therefore ignore these copayments in this paper.

fee-only patients.[11] The physician chooses a quality to offer the two groups. As we will argue below, the quality to the price-paying group will be set at the optimal level independent of fee, so we can normalize on this quality level, and regard the physician as choosing only the skimping to impose on the fee-only patients measured by s.

The physician can thus be regarded as making just two choices to maximize profit, the price for the price-paying patients, and the quality for the fee-only patients.[12] Once these variables are chosen, because the physician knows willingness to pay, and because services are non-retradable, the physician can charge a price to each patient (with its implicit quality discrimination) without fear that subsequent retrading will subvert the market segmentation.[13]

The physician's opportunity to supply services to other markets is represented in the cost function. To simplify the presentation, we assume marginal costs (for a given quality) are constant over quantity produced. Marginal cost for the price patients (with quality normalized at $s=0$) is c.[14] Skimping on quality lowers costs. We let $v(s)$ denote the cost saving per patient due to choice of s, with $v'>0$, and $v''<0$. We normalize the quality of services to the price-paying patients to $s=0$, and let $v(0)=0$. Cost per patient

[11] Patient willingness to pay appears to affect the likelihood of the physician asking for the extra price. In a regression explaining the probability of extra charge, patient income and supplemental insurance coverage were positively associated with the patient paying the extra price (Nelson et al., 1989). This empirical model did not separate, however, the effect of these variables on patients' choice of physician from the effect of physician's choice of charging a price.

A number of health economists argue that physicians have the ability to 'induce demand' in response to competition and fee changes. See, for example, Cromwell and Mitchell (1986), Dranove (1988), Rice and Labelle (1989) and McGuire and Pauly (1991) for discussion. In our model, physicians respond to fees by changing supply (which affects utilization) and by changing patients' demand through quality choice.

[12] We believe the assumption that physicians sort patients into two groups is only mildly restrictive. With regard to price discrimination, there will always be patients who are taken at the fee only. Our assumption of two groups is designed to capture the most important distinction in the market, between those who pay an additional price, and those who do not. A physician could get away with charging some patients in the price paying group more than others, but nothing near such perfect price discrimination is observed in practice. Rather than cloud the model with informational or transaction cost assumptions to justify a limited number of prices for the price-paying group, we assume the physician cannot perfectly price discriminate.

Similarly with respect to quality discrimination, there will always be patients who get the optimal quality. The most basic distinction is between these patients and those who get a degraded quality at the government-set fee. In principle a physician could get away with skimping more on some patients than others in the fee group, but this customized quality discrimination is so obviously unrealistic that we are comfortable with the assumption that quality for all in the fee-only group is the same.

[13] It is irrelevant for equilibrium that price-paying patients may prefer the price/quality package offered to fee-only patients, and vice versa. If the price-paying patients want services at the fee only, they have to seek another physician.

[14] The constant-cost assumption is innocuous, demand serving to limit the size of a physician's practice. In a working paper, we have shown that our results hold in the case of increasing marginal cost.

in the price market is therefore c. Variable s now denotes the difference between quality to the price patients and the fee patients. Cost per patient in the fee market is thus $c-v(s)$.

The assumption that quality to the price patients is normalized at zero deserves some discussion. If a patient pays a price p and the quality of service he gets is s, then his surplus from treatment is $\bar{U}-p-s-t$, where t is the distance from the physician. Notice that an increase of one dollar in price has the same effect on the patient's surplus as an increase in s by one. Therefore, if a patient is treated at a price p, it must be that the quality of service s he gets is such that $v'(s)=1$. If, for example, $v'(s)>1$, then the physician can decrease p by one and increase s by one. While the patient is indifferent to such a change and therefore will not change his decision, the physician is certainly better off since the decrease of revenue from the patient is one whereas the decrease in cost is $v'(s)>1$.

Thus it will always be the case that the service to the price-paying patients is set at a level s such that $v'(s)=1$, the efficient level of quality.[15] We can therefore normalize this level of quality to zero assuming that $v(0)=0$, and $v'(0)=1$. Variable s will then denote the difference between the quality to the price-paying patients and the quality to the fee-only patients.

2.3. *Market equilibrium*

It is important to stress that even though we introduce competition with only two physicians, our qualitative results will go through with more physicians. Clearly, if there are more physicians, the 'distance' between any two of them will be smaller. However, one can think of our model as studying the competition between physicians adjacent to one another in product space by normalizing the distance between them to be one. As long as we assume that some product differentiation exists, i.e., that any two physicians cannot locate exactly at the same point on the line, our results will hold.

We assume that all patients are served and let t_i^* denote the number of patients served by i (and hence, $1-t_i^*$ patients are served by j). For each of the t_i^* patients, physician i receives a fee f. The physician may price

[15]This result is sensitive to how quality is assumed to affect willingness to pay. Spence (1975) pointed out that the firm is concerned with the effect of quality on the marginal consumer, while social costs and benefits depend on the effects on the average consumer. We have assumed quality affects all patients equally, so the marginal patient (determining the physician's profit-maximizing quality choice) is affected the same as the average patient (determining the efficient choice). Other specifications of quality in demand would alter this result, although there would still be an efficiency gain from allowing a price to be charged if the monopolist tends to underproduce quality because the marginal patient's valuation is lower than the average patient's.

discriminate, so $\tilde{t}_i \leq t_i^*$ patients pay a price p_i, and $\hat{t}_i = t_i^* - \tilde{t}_i$ patients are treated for the fee only and pay nothing.

As we argued above, the quality of services to price-paying patients is constant. For physician i, s_i denotes the difference between the quality to patients served for the fee only and those who pay p_i, where s_i can be either positive (fee-only patients get lower quality) or negative (higher quality).

The number of patients willing to pay a given price to a physician depends on the patients' alternative: going to the other physician, receiving services of quality s_j, and paying nothing. Notice that the patients that are asked to pay the price p_i by physician i are those 'close' to physician i and hence 'far' from physician j. Therefore, if one of these patients goes instead to physician j, he will not be asked to pay p_j, but will be taken for the fee only. For a given physician's own price p_i and other's quality s_j, let $\tilde{t}_i$ satisfy the condition: $\tilde{t}_i + p_i = 1 - \tilde{t}_i + s_j$. Every patient at a distance from firm i less than or equal to $\tilde{t}_i$ prefers to pay p_i to physician i and get quality $s=0$ to paying zero to physician j and getting quality s_j.

Therefore,

$$\tilde{t}_i = (1 - p_i + s_j)/2 \tag{1}$$

is the number of patients that will agree to pay the price p_i to physician i rather than going for free to j.

We can now show how t_i^* is determined. t_i^* is the total number of patients served by physician i but it is also the distance from physician i of a patient who is just indifferent between being taken under the fee (paying nothing) by physician i receiving quality of treatment s_i and taken under the fee by physician j and getting quality of treatment s_j. Therefore, $t_i^* + s_i = 1 - t_i^* + s_j$. Any patient at a distance from physician i less than or equal to t_i^* prefers to be served under the fee by physician i than j.

Therefore,

$$t_i^* = (1 - s_i + s_j)/2 \tag{2}$$

is the total number of patients who are served by i out of which

$$\hat{t}_i = (p_i - s_i)/2 \tag{3}$$

are served at the fee only.[16]

[16]The nature of equilibrium will be affected by the degree of competitiveness of the two physicians – now 'close' they are in the geographical metaphor for competition used here. If the physicians are 'far apart,' so that in equilibrium $1 - t_i^* + s_j \leq \bar{U}$, physician j's choice of quality does not affect physician i's choice of price or quality, and each physician is a local monopoly. If physicians are closer competitors so that $1 - t_i^* + s_j \leq \bar{U}$, but $1 - \tilde{t}_i + s_j \geq \bar{U}$, competition affects choice of quality but not price. We choose to analyze the more representative case where any physician is likely to face some close competitors, and both price and quality-setting may be responsive to competitive pressures. We assume therefore that in equilibrium, $1 - t_i^* + s_j \leq \bar{U}$.

The profit to physician i is therefore,

$$\pi^i=(p_i+f-c)\tilde{t}_i+(f-c+v(s_i))\hat{t}_i. \tag{4}$$

Using (2) and (3), (4) can be expressed in terms of the physician's own strategy pair (p_i,s_i) and a strategy (p_j,s_j) of physician j:

$$\pi^i(p_i,s_i,p_j,s_j)=p_i(1-p_i+s_j)/2+(f-c)(1-s_i+s_j)/2+v(s_i)(p_i-s_i)/2. \tag{5}$$

A Nash equilibrium in this model is a pair of strategies such that (p_i^e,s_i^e) maximizes physician i's profit given (p_j^e,s_j^e). It must be, therefore, that (p_i^e,s_i^e) satisfies the following first-order conditions:

$$\partial\pi^i/\partial p_i=[1-2p_i^e+s_j^e+v(s_i^e)]/2=0 \tag{6}$$

$$\partial\pi^i/\partial s_i=[-(f-c)+v'(s_i^e)(p_i^e-s_i^e)-v(s_i^e)]/2=0. \tag{7}$$

and the following second-order condition:

$$-2v''(s_i^e)(p_i^e-s_i^e)+4v'(s_i^e)-(v'(s_i^e))^2\geq 0. \tag{8}$$

We examine the effect of fee in a symmetric equilibrium, letting $p_1^e-p_2^e=p$ and $s_1^e=s_2^e=s$. Conditions (6) and (7) become:

$$1-2p+s+v(s)=0 \tag{6'}$$

$$-(f-c)+v'(s)(p-s)-v(s)=0. \tag{7'}$$

From the first-order conditions (6) and (7) we also get that:

$$ds_i/ds_j=-v'(s_i)/[2v''(s_i)(p_i-s_i)+(v'(s_i))^2-4v'(s_i)].$$

By the second-order condition (8) we know that $ds_i/ds_j>0$, that is, the 'reaction functions' are upward sloping. We will confine our analyses to a stable equilibrium, which in our model implies that the reaction functions have a slope of less than 1 in equilibrium. Therefore, in the symmetric case:

$$v''(s)(1-s+v(s))+(v'(s))^2+3v'(s)<0. \tag{9}$$

This condition enables us to prove that quality increases uniformly with the fee. Substituting for p in (7') using (6'), we have:

$$v'(s)(1+v(s)-s)-2v(s)=2(f-c). \tag{10}$$

From (10),

$$ds/df = 2/[v''(s)(1+v(s)-s)+v'(s))^2+3v'(s)],$$

which must be negative by (9).

2.4. *Fees and equilibrium*

Physicians' equilibrium choice of quality and price depends on the level of fee set by the regulator. When the fee is low enough, no patients will be taken at the fee only. When the fee is high enough, no patients will be charged a balance bill. It is useful to define a fee, f^*, to be the fee that leads to an equilibrium quality choice of $s=0$ for fee patients. Denote by $\underline{f}$ the minimum fee necessary to induce physicians to take some patients at the fee only. We regard the range between $\underline{f}$ and f^* to be the 'normal' range. When the fee is set in that range some patients are served for the fee but the quality to the fee-only patients is less than or equal to the quality for the price patients. We show shortly that this is the range in which fee *should* be chosen. Appendix A to this paper details the full set of possible equilibria relating fees to prices and quality choices.

2.5. *Choosing the best fee*

We can use this model to characterize the optimal fee, first in the case in which price and quality discrimination is permitted. Since we focus only on symmetric equilibria in which patients are distributed evenly among physicians and since physicians' location is given, patients' transportation costs will not change as a result of a change in policy and, therefore, need not be considered. Also, since prices are only transfers from patients to physicians they do not enter into the welfare function.

Let r denote a particular patient and let R denote the set of all patients. Let $s(r)$ denote the quality of service that patient r receives. Total surplus can be measured by:

$$W=\int_{r\in R}(v(s(r))-s(r))dr-\theta f. \tag{11}$$

Recall that s measures the decrease in quality whereas $v(s)$ measures the savings in costs when quality is decreased. Therefore, $v(s)-s$ measures the net 'gain' to society when quality is decreased. Notice though that $v(s)-s$ equals zero when s is zero and that $v(s)-s<0$ for any other s. Therefore, from society's point of view, it would be optimal if all patients received quality $s=0$.

The second term in (11), $-f\theta$, measures the cost to society as a result of the distortion created by the tax financing f.[17] We assume that $\theta>0$.

[17] If the payer were a private insurer, θ would measure administrative costs, typically about 15 percent of gross payouts.

Therefore, if a planner could dictate quality, he would set $s=0$ and $f=0$. If, however, quality cannot be dictated, the planner can only induce quality by choosing the fee. For each level of fee the planner knows what will be the equilibrium quality, and so he will choose fee so as to induce the quality that will maximize surplus. As we will show below, with distortionary finance, the second-best quality will involve $s>0$.

Let us now focus on the welfare properties of the equilibrium. Patients who pay the price receive the optimal quality. All other patients receive the same quality s which, as equation (A5) from Appendix A shows, is a function of f and can therefore be denoted by $s(f)$. For a given f, the number of patients who are taken at the fee by both physicians is $(1+v(s(f))-s(f))/2$, in equilibrium. Let $W^p(f)$ (for 'welfare with price discrimination') specify the welfare, in equilibrium, when the fee level is f. Then,

$$W^p(f)=(v(s(f))-s(f))(1+v(s(f))-s(f))/2-\theta f. \tag{12}$$

Let f^p be the 'constrained' optimal level of fee, i.e., the level of fee at which equilibrium surplus is maximized. Then we can state the following result about the optimal fee: $s(f^p)>0$, i.e., at the 'constrained' optimal level of fee, quality is 'too low' relative to the first-best optimum. Formally, notice that f^* is defined such that $s(f^*)=0$. It is easily checked, however, that $dW^p(f^*)/df=-\theta<0$. Thus, it must be that $f^p<f^*$ which implies that $s(f^p)>0$.

To see why these results hold, consider raising the fee towards the level which leads to an equilibrium with $s=0$, the first-best quality. The social cost of a higher fee in terms of distortionary finance is constant and proportional to the level of the fee. The social gains in terms of the efficient level of quality $(v(s(f))-s(f))$ are decreasing as the fee is raised. The second-best (constrained) optimum is where the marginal gain, still positive, is just equal to the social cost. This must be at a fee leading to an $s>0$. Distortionary tax financing implies that the fee should be kept below a level that equalizes the quality to the price paying patients and fee patients. Only if the payer could raise funds costlessly, should f^p equal f^*.

Before going on to consider the effects of balance billing, it is worth noting another implication of the model. If the physician accepts any patients at the fee ($\hat{t}>0$), and quality is variable, fee exceeds marginal cost. This feature of equilibrium is a result of physician profit maximization with respect to quality. If marginal cost equals the fee, the physician could reduce quality for the fee patients and increase profit. The quality reduction would decrease cost for all patients, whereas the loss of the marginal patient would have only a second-order impact on profits. (Equilibrium could obviously never take place when marginal cost exceeds the fee – the physician would decline patients.) This observation is an explanation for Pauly's (1991) contention

that fees are 'too high,' exceeding cost, and physician services in Medicare are in 'excess supply.' Recognizing that physicians exercise market power in the fee market by setting quality and cost, the condition that fees exceed marginal cost cannot be eliminated even if fees are reduced. Thus it cannot be concluded that just because fees exceed marginal cost, fees should be reduced. Choice of the optimal fee must take into account the effect on quality.

3. Should physicians be permitted to 'balance bill'?

In addition to deciding the level of the fee, a payer decides if the fee must be taken as full payment. Policy towards 'balance billing,' has been continually revised by Medicare since the introduction of the 'participating physician program,' first discouraging doctors from price discriminating. Gradually, price discrimination has been punished more, and limits have been made more severe on permissible price discrimination.[18] In this section, we consider the effects of a policy that bans price discrimination altogether. When price discrimination is banned, we also do not permit any quality discrimination since there is no longer a natural separation of patient markets. Thus, in this section, physicians must offer the same quality to all patients and accept all patients for the fee only.

We first solve for equilibrium with no discrimination. Let t_i denote the number of patients served by physician i. Notice that t_i is the distance from physician i of the patient who is just indifferent between being taken at the fee by physician i (getting quality s_i) and taken at the fee by physician j (at quality s_j). Then,

$$t_i=(1+s_j-s_i)/2$$

and the physician's profit is

$$\Pi i(s_i,s_j)=(f-c+v(s_i))(1+s_j-s_i)/2. \tag{13}$$

The first term in (13) is the profit per patient for physician i, and the second is the number of patients served by him. The first-order conditions imply that in the symmetric case:

$$v'(s)-(f-c+v(s))=0. \tag{14}$$

Determination of quality without discrimination from (14) can be compared with the quality from (B5) in Appendix B. Our finding is stated as a proposition (proved in Appendix B):

[18]In 1991, price was limited to 25 percent of the fee applicable to non-participating physicians. This fell to 15 percent in 1993.

Proposition: For a given fee, when balance billing and quality discrimination are prohibited, quality to those who are served under the fee is lower than with discrimination.

The reason is that with discrimination prohibited, physicians can only extract rents by setting quality. They do so by reducing quality, and therefore saving on costs.

This finding can be used to compare welfare under the optimal fee when balance billing is prohibited to welfare under the optimal fee when balance billing is permitted. As we will show, the latter is always higher than the former.

For every f let $s(f)$ be the level of quality that satisfies the equilibrium condition (14). Therefore, for every f welfare in equilibrium when balance billing is prohibited is simply:

$$W^o(f)=v(s(f))-s(f)-\theta f. \tag{15}$$

Notice that here all patients get the same quality and the number of patients is one. $W^o(f)$ is maximized at f^o such that:

$$\partial W^o/\partial f=[v'(s(f^o))-1]\mathrm{d}s/\mathrm{d}f-\theta=0. \tag{16}$$

Proposition: Allowing price and quality discrimination, there is always a fee that leads to a higher welfare than can be achieved if price and quality discrimination were prohibited, i.e., $W^p(f^p)>W^o(f^o)$.

It should not be surprising that overall welfare is increased by allowing firms with market power to price discriminate. We give the intuition for this result here, and the proof in Appendix B. Suppose that initially balance billing is prohibited and fee is set at the optimal level, f^o. Suppose now balance billing and quality discrimination are permitted and fee is kept at f^o. Social surplus on those patients who are charged a price clearly increases, since they now get the optimal level of quality. What happens to those patients who are now taken at the fee? We know that quality for these patients goes up (i.e., s goes down). This, by itself does not mean that social surplus on these patients is increased since it is possible that quality is improved 'too much' (i.e., $v(s)-s$ is decreased). However, if this is the case, the planner can always reduce the fee. This will not only reduce the quality for the fee-only patients toward the socially optimal level (without disturbing the quality offered to the price-paying group), but it will also reduce the cost of distortionary finance. Hence, surplus will go up.

4. Payer fee discrimination

In the analysis so far, we have required the payer to set the same fee for all

patients. When the physician discriminates and charges price to one segment of the market, should the payer set the same fee for the price and fee-only patients? We consider in this section an incremental and a drastic version of 'fee discrimination' by the payer.

The first is a small amount of payer fee discrimination, paying a slightly higher fee for patients if the physician agrees to accept the fee as payment in full. In practice this would mean if a claim is taken on assignment (no balance bill charged), Medicare would pay the usual fee. If the claim is not assigned, the physician's fee would be reduced by some amount, but balance billing would still be allowed. Medicare currently practices a different form of fee discrimination by paying some physicians, those who agree to treat all patients for the fee, at slightly higher fees. Our suggested policy is one that discriminates *within* the practice of a single physician rather than *across* the practices of different physicians.[19]

Consider, then, the following policy: A physician is paid a fee of $f+d$ if he takes a patient under the fee and he is paid a fee of only $f-d$ if he charges the patient a price of $p>0$.

Proposition: When d is small and f is near f^p, a small increase in d will increase surplus.

A small amount of fee discrimination has two countervailing effects on welfare. On one hand it increases quality to the fee patients which brings them closer to the socially optimal quality level. On the other hand, as a result of the fee discrimination, some patients who were served under the price (when fee discrimination was not present i.e., $d=0$) and received the socially optimal quality will be now taken under the fee and will therefore receive a lower quality. We prove in Appendix B, however, that the first effect is dominant when the fee is set at near the optimal level. Thus, when d is small and f is near f^p, a small increase in d will increase surplus. A small amount of payer fee discrimination can certainly increase welfare.

The second policy we consider is a more drastic form of payer fee discrimination: it is the complete elimination of the fee for patients the physician charges a price. The physician could continue to accept patients for the fee, but there would be no fee for the price-paying patients.

While this drastic form of fee discrimination may seem extreme to an American audience, it actually closely mirrors the national health policy of most other countries. Physicians in much of the world allocate their work time working between public sector patients seen for a fee, and private patients charged a price. In England, Israel, India, and many other countries,

[19]Non-participating physicians are paid 95 percent of the fee schedule amount. This small differential is to be maintained in existing legislation.

physicians operate a public and private practice, and receive no public contribution if they see a patient privately.

Proposition: There exists θ^* such that if $\theta > \theta^*$ ($\theta < \theta^*$) the policy of paying fee only for patients that do not pay the price will increase (decrease) welfare relative to the policy of paying the same fee for all patients.

In examining this policy we find that it may be optimal to drive the fee for the price-paying patients to zero, but only if the shadow price of public funds, θ, is sufficiently high. In equilibrium under this fee policy, more patients are seen for the fee only. However, fee patients receive a lower quality than they would without payer fee discrimination. This is because with more patients seen at the fee, the physician faces a larger return from a quality decrease to the fee patients. (See Appendix B.)

5. Conclusion

The federal Medicare program is committed to a fee restructuring, and other public and private payers can be expected to follow suit. Under the new fee policy, the Medicare program will set the overall level of fees, and the terms on which a physician can charge a supplemental price, or 'balance bill.' This paper has analyzed the economic effects of fee policies in a model where physicians are not homogeneous, and may price and quality discriminate in competition for patients.

Our findings have some immediate implications for Medicare and other payers setting a fee policy. First, in equilibrium, there will be no excess demand and fees will exceed marginal cost. This will be true whether fees are currently 'too high' or 'too low.' A difficult judgment about quality is necessary in deciding the right level of fees.

Our second conclusion bears on the controversial issue of balance billing. We show that there are efficiency benefits from permitting price discrimination in the form of balance billing. When price discrimination is permitted, quality is set at a higher level for both patients paying the price *and* those not paying a supplemental price. Efficiency of equilibrium is improved. This strong efficiency result must be weighed against distributional considerations in any policy choice.

Finally, we recommend that fee policy include an element of payer fee discrimination, wherein a physician would be paid more on behalf of patients not charged a price than for those taken only for a fee. This is a mild version of what national health payment systems do in much of the world, where a physician may operate a private practice, but receives no government payment for a patient seen privately. At least a small step in that direction is warranted in the U.S., in the form of a deduction from the normal fee when the physician chooses to charge a price.

Appendix A

This appendix characterizes equilibrium when two physicians can price and quality discriminate. The nature of equilibrium depends on the level of the fee set by the regulator.

To describe the effect of fee on the nature of equilibrium, it is useful to define critical values of the quality variable s. These values of s will correspond to critical values of f in the proof below. From (6'), price will be set at or above zero when:

$$1+s+v(s)\geq 0. \tag{A1}$$

Condition (A1) sets a lower bound for s. Notice next that $(p-s)/2$ is the number of patients served by each physician in equilibrium under the fee only. For this to be positive, also from (6'):

$$1-s+v(s)\geq 0. \tag{A2}$$

This sets an upper bound for s. Together, conditions (A1) and (A2) imply that when equilibrium has both price discrimination and patients taken for the fee, s must fall in the following range: $\underline{s}\leq s\leq \bar{s}$, where $\underline{s}<0$ satisfies

$$1+\underline{s}+v(\underline{s})=0 \tag{A3}$$

and $\bar{s}>0$ satisfies

$$1-\bar{s}+v(\bar{s})=0. \tag{A4}$$

We can now present the following summary of the relation between quality and the fee in equilibrium:

Proposition: There exist fee levels $\underline{f}, f^*$, *and* $\bar{f}$ *such that* $\underline{f}<f^*<\bar{f}$ *and*

(*a*) *if* $f<\underline{f}$ *no patient is served under the fee.*

(*b*) *if* $\underline{f}<f<f^*$ *quality to patients under the fee is lower than quality to those who are served under the price* (*i.e.*, $s>0$)

(*c*) *if* $f^*<f<\bar{f}$ *quality to those who are served under the fee is higher than the quality to those who pay the price* (*i.e.*, $s<0$).

(*d*) *if* $f>\bar{f}$ *there is no price discrimination* (*i.e.* $p=0$), *but there is quality discrimination.*

Proof: Substituting for p in (7') using (6'), we have:

$$v'(s)(1+v(s)-s)-2v(s)=2(f-c). \tag{A5}$$

Let $\underline{f}$ be the fee for which the equilibrium quality is $\bar{s}$. Then, using the definition of $\bar{s}$, (A4), $\underline{f}$ must satisfy $-2v(\bar{s})=2(\underline{f}-c)$, which implies that

$$\underline{f} = c - v(\bar{s}). \tag{A6}$$

Let f^* be the fee for which the equilibrium quality is $s=0$. Then, by (A5), f^* satisfies:

$$1 = 2(f^* - c)$$

which implies that

$$f^* = c + .5. \tag{A7}$$

Since $v(\bar{s}) > 0$, (A6) and (A7) imply that $f^* > \underline{f}$.

Let $\bar{f}$ be the fee for which the equilibrium quality is $\underline{s}$. Then, by (A5) and (A3), $\bar{f}$ satisfies:

$$v'(\underline{s})(-2\underline{s}) - 2v(\underline{s}) = 2(\bar{f} - c)$$

which implies that

$$\bar{f} = c - (\underline{s}v'(\underline{s}) + v(\underline{s})). \tag{A8}$$

Since $v'(\underline{s}) > 1$ and since $-(s + v(\underline{s})) = 1$ it must be that $\bar{f} > f^*$.

We shall now show that if $f < \underline{f}$, no one is served under the fee. Observe first that by the stability condition (9), the left-hand side (lhs) of (A5) is decreasing with s. If f were less than $\underline{f}$, s would have to be above $\bar{s}$ for condition (A5) to hold. When $s > \bar{s}$, no patients are served at the fee.

In a similar way it can be shown that if $f > \bar{f}$, $s = \underline{s}$ which implies that the price is zero. This result follows from the existence of competition among physicians.

If $\underline{f} < f < f^*$ the monotonicity of the lhs of (A5) implies that $s > 0$ and if $\bar{f} > f > f^*$ it must be that $s < 0$. Furthermore, using the stability condition (9), it can be shown that throughout the range $\underline{f} \leq f \leq \bar{f}$, $ds/df < 0$. Increases in fee beyond this point continue to increase quality. This completes the proof.

Appendix B

This appendix states and proves propositions mentioned in the text.

Proposition: For a given fee, when balance billing and quality discrimination are prohibited, quality to those who are served under the fee is lower than with discrimination. Furthermore, for those who pay the price, quality is lower (higher) if $f - c \leq 1$ (≥ 1).

Proof: Notice that the equilibrium condition (14) can be rewritten as:

$$v'(s) - v(s) = f - c. \tag{B1}$$

Clearly s is negative (positive) if the right-hand side is greater (smaller) than 1. Also notice that since $v'(s) > v'(s)(1-v(s)-s)/2$ for all $\underline{s} \leq s \leq \bar{s}$, our first conclusion holds.

Proposition: Allowing price and quality discrimination, there is always a fee that leads to a higher welfare than can be achieved if price and quality discrimination were prohibited, i.e., $W^p(f^p) > W^o(f^o)$.

Proof: We first establish the following lemma.

Lemma: If $\theta > 0$, $f^o - c < 1$.

Recalling (16), f^o is defined as the solution to

$$\partial W^o/\partial f = [v'(s(f^o)) - 1]\mathrm{d}s/\mathrm{d}f - \theta = 0.$$

From (A1) it is easy to see that $\mathrm{d}s/\mathrm{d}f < 0$ and that $v'(s(f)) - 1 < 0$ if and only if $f - c < 1$.

Using this Lemma we can now prove the proposition. Let $s^p(f)$ be the quality to the fee patients when discrimination is not prohibited and the level of fee is f and let $s^o(f)$ be the quality to *all* patients when discrimination is prohibited.

Suppose that $f = f^o$ so that fee is set at the optimal level if discrimination is prohibited. From the previous Proposition 4 we know that $s^p(f^o) < s^o(f^o)$.

It is also clear that $s^o(f^o) > 0$ since otherwise welfare, in the no discrimination case, can be increased by reducing the fee.

Thus it follows that if at f^o we allow discrimination to take place, quality to both the price patients and the fee patients will be higher.

If $s^o(f^o) > s^p(f^o) > 0$, then it is clear that $W^p(f^p) > W^p(f^o) > W^o(f^o)$. If, on the other hand, $s^p(f^o) < 0$, then quality to the fee patient gets 'too high' when discrimination is permitted. In this case, however, we can decrease the fee which will in turn decrease quality and will increase welfare. This completes the proof.

Proposition: When d is small and f is near f^p, a small increase in d will increase surplus.

Proof: Notice first that

$$\mathrm{d}s/\mathrm{d}d = -(2v''(s) - 1)/(\mathrm{d}A/\mathrm{d}s) < 0, \tag{B2}$$

where $A = v'(s)(1 + v(s) + 2d - s) - 2v'(s)$.

Also, $\hat{t}$, the number of patients taken under the fee by each physician, is

$$\hat{t} = (1 + v(s) + 2d - s)/4. \tag{B3}$$

Therefore,

$$d\hat{t}/dd = [2 + (v'(s) - 1)(ds/dd)]/4 > 0. \quad \text{(B4)}$$

Now, for a given d, total surplus is equal to:

$$W(d) = (v(s(d)) - s(d))2\hat{t}(d) - \theta[2\hat{t}(d)(f + d) + (1 - 2\hat{t}(d))(f - d)]. \quad \text{(B5)}$$

The first term in $W(d)$ measures the effect of quality of treatment of the fee patients on welfare, whereas the second term measures the cost associated with the financing of the fee. Differentiating $W(d)$ with respect to d we get:

$$dW(d)/dd = (v'(s) - 1)(ds/dd)2\hat{t}(d) + 2(d\hat{t}/dd)(v(s) - s) - \theta[4\hat{t} - 1 + 4(d\hat{t}/dd)d]. \quad \text{(B6)}$$

Equation (B6) has three terms. The first term is positive (for $s > 0$) and the second one is negative. For $s > 0$ we already know that $\hat{t} < 1/4$ and since $d\hat{t}/dd$ is continuous, it is clear that the third term in (B6) is positive for a small enough d. Since the third term is positive it is enough to show that the first term is bigger, in absolute value, than the second one in order for (B6) to be positive. Dividing the second term by the first one we get:

$$(d\hat{t}/dd)(v(s) - s)/[(v'(s) - 1)(ds/dd)t]. \quad \text{(B7)}$$

Using l'Hôpital's rule we can show that (B7) converges to zero as s converges to zero and, hence, the first term in (B6) is larger than the second one (in absolute value) and hence (B6) is positive. This completes the proof.

Proposition: There exists θ^* such that if $\theta > \theta^*$ ($\theta < \theta^*$) the policy of paying fee only for patients that do not pay the price will increase (decrease) welfare relative to the policy of paying the same fee for all patients.

Proof: The policy above has two countervailing effects on welfare. On one hand it decreases quality and hence welfare. On the other hand since fee is paid only for the fee patients, the distortion associated with financing the fee is reduced. Clearly for a θ large enough the second effect is dominant.

References

Baumol, W.J., 1988, Price controls for medical services and the medical needs of the nation's elderly, March 11.

Borenstein, S., 1985, Price discrimination in free entry markets, Rand Journal of Economics 16, 380–397.

Cromwell, J. and J. Mitchell, 1986, Physician-induced demand for surgery, Journal of Health Economics 5, 293–313.

Cromwell, J., J. Mitchell, M. Rosenbach, W. Stason and S. Hurdle, 1989, Using relative physician effort to identify mispriced procedures, Inquiry 26(1), 7–23.

Dranove, D., 1988, Demand inducement and the physician/patient relationship, Economic Inquiry 26, 281–298.
Feldman, R. and F. Sloan, 1988, Competition among physicians, revisited, Journal of Health Politics, Policy and Law 13, 239–261.
General Accounting Office, 1989, Impact of state mandatory assignment programs on Beneficiaries, GAO/HRD-89–128, September (Washington, DC).
Holmes, T.J., 1989, The effects of third-degree price discrimination in oligopoly, American Economic Review 79(1), 244–250.
Hsiao, W.C., P. Braun, E. Becker, N. Causino et al., 1988, A national study of resource-Based relative values for physician services: final report (Harvard School of Public Health) September 22.
Katz, M. 1984, Price discrimination and monopolistic competition, Econometrica 53, 1453–1472.
Lederer, P.J. and A.P. Hurter, Jr., 1986, Competition of firms; discriminatory pricing and location, Econometrica 54, 623–640.
Maskin, E. and J. Riley, 1984, Monopoly with incomplete information, Rand Journal of Economics 15, 171–196.
McGuire, T.G. and M.V. Pauly, 1991, Physician response to fee changes with multiple payers, Journal of Health Economics 10, 385–410.
Mitchell, J. and J. Cromwell, 1982, Physician behavior under the Medicare assignment option, Journal of Health Economics 1, 245–264.
Nelson, L., A. Ciemnecki, N. Carlton and K. Langwell, 1989, Assignment and the participating physician program: an analysis of beneficiary awareness, understanding and experience, Mathematica Policy Research Inc., MPR Ref. No.: 7815–700.
Pauly, M., 1991, Fee schedules and utilization, in: H.E. Frech III, ed., Regulating doctors' fees: competition, benefits, and controls under Medicare (AEI Press, Washington, DC).
Physician Payment Review Commission, 1990, Annual Report to Congress, Author (Washington, DC).
Prospective Payment Assessment Commission (ProPAC), 1989, Medicare prospective payment and the American health care system: Report to the Congress, June (Washington, DC).
Rice, T.H. and R.J. Labelle, 1989, Do physicians induce demand for medical services?, Journal of Health Politics, Policy and Law 14(3), 587–601.
Sheingold, S.H., 1989, The first three years of PPS: impact on Medicare costs, Health Affairs 8(3), 191–204.
Spence, A. M., 1975, Monopoly, quality and regulation, The Bell Journal of Economics 6(2), 417–429.
Tirole, J., 1988, The theory of industrial organization (The MIT Press, Cambridge, MA).
Wedig, G., J.B. Mitchell and J. Cromwell, 1989, Can optimal physician behavior be obtained using price controls? Journal of Health Politics, Policy and Law 14(3), 601–620.
White, L.J., 1981, Reforming regulation: processes and problems (Prentice Hall, Englewood Cliffs, NJ).
Zuckerman, S. and J. Holahan, 1991, The role of balance billing in Medicare physician payment reform, in: H.E. Frech III, ed., Regulating doctors' fees: competition, benefits, and controls under Medicare (AEI Press, Washington, DC).