THE UNIVERSITY OF SYDNEY

Economics Working Paper Series

2012 - 12

An Equilibrium Model of General Practitioner Payment Schemes

Donald J. Wright

July 2012

An Equilibrium Model of General Practitioner Payment Schemes

Donald J. Wright*

July 2012

Abstract

In an environment where GPs are of differing quality and heterogeneous patients have different preferences for quality, it is shown that fee-for-service coupled with balance billing is a superior payment scheme to just fee-for-service or capitation payments as it generates an efficient allocation of GPs between high and low quality and an efficient allocation of patients between GPs. Where patients have more than one condition it is shown that fee-for-service allows patients to seek treatment from GPs of differing quality conditional on the medical condition they have.

JEL Classification Numbers: I11

Keywords: General Practitioner Payment Schemes

*School of Economics, Faculty of Arts and Social Sciences, University of Sydney, NSW, 2006, Australia, Ph: 61+2+93516609, Fax: 61+2+93514341, Email: don.wright@sydney.edu.au.

1 Introduction

The basic problem faced by a government purchaser of primary health-care is how to induce general practitioners to provide the welfare maximising quantity and quality of health-care to patients. The difficulty is that the quality of care is not observed by the purchaser and so general practitioner (GP) payment schemes can not be made contingent on quality, they can only be made contingent on observed quantity. However, quality of care might be observed by patients. In fact, much of the literature assumes just this, Glazer and McGuire (1993), Ellis (1998), Gravelle and Masiero (2000) and Karlsson (2007). In the first two of these papers quality is known by patients, in the last two papers the probability distribution of quality is known in a first period and this probability is updated in a second period. Ma and McGuire (1997) argue that because GPs have long-term relationships with patients it is reasonable to assume that patients actually observe GP quality.

The fact that GP quality is observed by patients allows the purchaser some leverage over quality through its choice of GP payment scheme. In the papers by Ellis (1998), Gravelle and Masiero (2000), and Karlsson (2007), a government purchaser chooses a capitation payment (a payment per registered patient) and GPs compete for patients through their choice of quality.[1] The emphasis is on whether or not a capitation payment provides GPs with the appropriate incentives for efficient quality choice. Patients in these papers value quality identically and differ by their location in geographic space.

Glazer and McGuire (1993) also have GPs competing for patients by choosing quality. The purchaser pays GPs by fee-for-service. In addition to

[1] Strictly speaking Ellis (1998) is concerned with hospital payment schemes but the analysis carries over to GPs. He also includes some cost reimbursement in the payment scheme.

this payment from the purchaser, GPs can balance bill by charging patients a fee-for-service directly. They showed that allowing balance billing increases welfare because it allows GPs to discriminate by offering a high quality service to patients that are balanced billed and a low quality service to patients that are not balance billed.

In this paper, patients are assumed to know GP quality but have different preferences for quality. The emphasis is not on whether a payment scheme induces GPs to choose the efficient quality but rather whether it induces an efficient allocation of GPs between high and low quality and whether it induces an efficient allocation of patients between GPs. To the authors knowledge, allocation questions of this kind have not been addressed in the GP payment scheme literature.

To analyse allocation issues a relatively simple model is developed. GPs, are assumed to value income and the quality of the service they provide. This quality of service can be either high or low and is chosen by GPs. The total number of GPs is fixed. There is a fixed number of patients who value quality differently and choose which GP to visit. Two payments schemes are considered, fee-for-service and capitation. Under fee-for-service a government purchaser pays a fixed price for all GP services and also determines whether GPs can balance bill.[2]. Under capitation, patients register with a GP and receive all their primary health-care from the GP they register with. The government purchaser pays GPs a fixed amount per-patient and also determines whether GPs can charge a fee-for-service directly, that is,

[2]Balance billing was used in the US for Medicare patients in the early 1980's and is the current payment scheme in Australia, where GPs can 'bulk bill' patients (patients pay zero and GPs receive a fixed payment from the government for each service delivered) or can choose to not 'bulk bill' (patients pay a price greater than the fixed payment but get reimbursed the amount of the fixed payment)

balance bill.

The main result of this paper is that under fee-for-service allowing balance billing result in an efficient allocation of GPs between high and low quality and also an efficient allocation of patients between GPs. This is not possible without balance billing. Essentially, an optimally chosen purchaser paid fee-for-service coupled with a patient paid fee-for-service, which is determined by the forces of supply and demand, duplicates the outcome of a competitive equilibrium. Patients who value quality more highly are treated by high quality GPs and pay a fee for their services while patients who value quality less highly are treated by low quality GPs and pay nothing for their services. It is also shown that markets in which patients have a stronger preference for high quality on average, have more high quality GPs and more patients allocated to high quality GPs. In addition, it is shown that where patients have one of two conditions, and one of these conditions is such that the quality of the GP is unimportant, then patients with a relatively high preference for quality choose to be treated by a low quality GP for the condition for which quality is unimportant and for the other condition they choose to be treated by a high quality GP and pay an additional fee. In all of these cases fee-for-service with balance billing maximises welfare.

Under payment by capitation it is shown that capitation and balance billing yields the efficient allocation of GPs between high and low quality and also the efficient allocation of patients between GPs if patients have one condition. However, where patients have one of two conditions and for one of those conditions quality is unimportant, then a capitation payment can not yield the welfare maximum as the welfare maximum involves some patients seeing high quality GPs for one condition and low quality GPs for

the other. This is not possible under a capitation payment where patients register with a particular GP and receive all their health-care from that GP. Therefore, an optimally chosen fee-for-service coupled with balance billing is a superior payment mechanism to capitation, even with balance billing, as it allows the movement of patients between GPs depending on the condition they have.

It should be noted that Glazer and McGuire (1993) demonstrate that fee-for-service coupled with balance billing is superior to just fee-for-service. This is similar to the main result of this paper. However, in their paper, GPs offer a different quality of service to different patients while in this paper GPs offer the same quality of service to all patients but some GPs are high quality and others are low quality. In Australia, GPs tend to either bulk bill all patients (no balance billing) or balance bill all patients. This is consistent with GPs offering the same quality of service to all patients they serve and not consistent with the model of Glazer and McGuire.

2 General Practitioners, Patients, and the Purchaser under Fee-for-Service

2.1 GPs

GPs derive utility from income, y, and also the quality of the service they provide. This quality of service can be either high quality, $q_H > 0$, or low quality, $q_L > 0$. The utility function of a GP with preference parameter γ is

$$U(y;\gamma) = \gamma \cdot q_H + y. \tag{1}$$

Providing a low quality service provides no utility to any GP. The preference parameters γ is distributed on the interval $\gamma \in (0,1]$ with density $t(\gamma)$ and cumulative distribution $T(\gamma)$. The total number of GPs is fixed at $\bar{G}$.

Under fee-for-service GPs get paid a price p for each service provided and have a cost function which is an increasing convex function of the number of services provided, n, and also depends on the quality of service. This cost function is given by $c_i(n),\ i = H, L$, where $c_i'(n) > 0$, $c_i''(n) > 0$, and $c_H'(n) > c_L'(n)$. The last inequality states that marginal cost is greater for the high quality service than the low quality service. Perhaps a high quality service requires more GP time than a low quality service. GP income is $y = p \cdot n - c_i(n)$.

2.2 Patients

Patients are assumed to have Mussa and Rosen (1978) preferences, so a patient with preference parameter θ obtains surplus

$$V = \theta \cdot q_i - p_p \tag{2}$$

when purchasing a GP service of quality q_i at a price of p_p, and zero otherwise. The individual preference parameter, $\theta > 0$, is distributed on the interval $[\underline{\theta}, \overline{\theta}]$ with density $f(\theta)$ and cumulative distribution $F(\theta)$.

It is assumed that there are $\bar{N}$ patients who have the same condition and have unit demands for GP services. The number of patients with preference parameter greater than θ is $N(\theta) = (1 - F(\theta)) \cdot \bar{N}$. It is also assumed that patients can observe GP quality.[3]

2.3 Purchaser

It is assumed that GP services are purchased on behalf of patients by a government purchaser. The government purchaser pays the GP p_g per-

[3]This assumption is standard in the literature in which GPs (Hospitals) compete for patients through their choice of quality, Glazer and McGuire(1993), Ma and McGuire (1997), Ellis (1998), Gravelle and Masiero (2000), and Karlsson (2007).

service. Therefore, the price the GP receives is $p = p_p + p_g$. If $p_p = 0$ the patient has complete insurance. The case where $p_g > 0$ and $p_p > 0$ is referred to as balance billing in the United States. In Australia, under Medicare, $p_p > 0$ is referred to as an out-of-pocket expense.

3 Welfare Maximum

As a point of comparison it is useful to solve for the welfare maximising number of high and low quality GPs, G_H and $G_L = \bar{G} - G_H$, respectively, and the welfare maximising allocation of patients between these GPs, N_H and $N_L = \bar{N} - N_H$. Let the inverse of $N(\theta)$ be $\theta(N)$, where $\theta'(N) < 0$. $\theta(N)$ is the θ which has N patients with preference parameter greater than or equal to θ. It can also be interpreted as the preference parameter of patient N, where patients have been ordered from highest θ to lowest θ. Let the N_H patients with the highest $\theta's$ be allocated to high quality GPs, that is, patients with preference parameters θ in $[\theta(N_H), \overline{\theta}]$. The remaining patients are allocated to low quality GPs, that is, those patients with parameters θ in $[\underline{\theta}, \theta(N_H))$. To minimise cost, patients are allocated between GPs of the same quality type so that marginal costs are equalized. Given all GP's have the same cost function, the cost of serving N_i patients by G_i GPs of quality i is $c_i(\frac{N_i}{G_i}) \cdot G_i$.

Let GPs be ordered according to their preference parameter and let the preference parameter of GP G be given by $\gamma(G)$, where $\gamma'(G) < 0$. That is, GPs with the highest preference parameters are ordered first. Welfare is

given by,

$$\begin{aligned} W &= \int_0^{N_H} \theta(N) \cdot q_H \cdot dN - c_H\big(\frac{N_H}{G_H}\big) \cdot G_H \\ &+ \int_{N_H}^{\bar{N}} \theta(N) \cdot q_L \cdot dN - c_L\big(\frac{\bar{N} - N_H}{\bar{G} - G_H}\big) \cdot (\bar{G} - G_H) \\ &+ \int_0^{G_H} \gamma(G) \cdot q_H \cdot dG. \end{aligned} \tag{3}$$

Assumptions: The following assumptions are made to ensure the welfare maximum involves all patients being served and that at least one patient is served by a low quality GP and at least one is served by a high quality GP.

$$\underline{\theta} \cdot q_L > c'_L\big(\frac{\bar{N}}{\bar{G}}\big) \tag{4}$$

This states that the utility of the patient that values quality the lowest is greater than the marginal cost of serving this patient if all GPs are low quality.

$$\overline{\theta} \cdot q_H - c'_H(1) > \underline{\theta} \cdot q_L - c'_L(\frac{\bar{N}}{\bar{G}}) \tag{5}$$

$$\underline{\theta} \cdot q_L - c'_L(1) > \underline{\theta} \cdot q_H - c'_H(\frac{\bar{N}}{\bar{G}}) \tag{6}$$

(5) states that the marginal net benefit of treating one patient by a high quality GP is greater than the marginal net benefit of treating all patients with low quality GPs and (6) states that the marginal net benefit of treating one patient by a low quality GP is greater than the marginal net benefit of treating all patients with high quality GPs

Welfare is maximised by choosing the number of high quality GPs, G_H, and the allocation of patients to high quality GPs, N_H. Differentiating welfare with respect to N_H and G_H yields the following first order conditions

for a maximum

$$\theta(N_H)\cdot(q_H-q_L)=c'_H\Big(\frac{N_H}{G_H}\Big)-c'_L\Big(\frac{\bar{N}-N_H}{\bar{G}-G_H}\Big) \tag{7}$$

and

$$c'_L(\cdot)\cdot\Big(\frac{\bar{N}-N_H}{\bar{G}-G_H}\Big)-c_L\Big(\frac{\bar{N}-N_H}{\bar{G}-G_H}\Big)=c'_H(\cdot)\cdot\Big(\frac{N_H}{G_H}\Big)-c_H(\cdot)+\gamma(G_H)\cdot q_H. \tag{8}$$

The second order conditions for a maximum are given in the Appendix and hold by assumption. Let the solutions to these first order conditions be unique and denoted by N_H^* and G_H^*. Patients with $\theta<\theta(N_H^*)$ are treated by low quality GPs and patients with $\theta\geq\theta(N_H^*)$ are treated by high quality GPs.[4] GPs with preference parameter $\gamma\geq\gamma(G_H^*)$ are high quality and GPs with $\gamma<\gamma(G_H^*)$ are low quality.

Given G_H^*, condition (7) allocates patients between GPs so that the extra marginal benefit of having a patient with preference parameter $\theta(N_H^*)$ treated by a high quality GP equals the extra marginal cost of doing so. Similarly, given $\theta(N_H^*)$, condition (8) allocates GPs between high and low quality so that the marginal cost of having the G_H^{th} GP be low quality equals the marginal cost of having this GP be high quality adjusted for the preference this GP has for delivering a high quality service. Essentially, G_H^* minimises the net cost of having N_H^* patients being treated by high quality GPs.

[4] Tirole (1988, p 96-97) provides a reinterpretation of the preferences given in (2) above, where consumers have identical preferences but differ in their incomes. In this reinterpretation, patients with higher incomes have lower marginal utilities of income and higher θ. Therefore, in the welfare maximum, it is high income patients that are treated by high quality GPs and low income patients that are treated by low quality GPs.

4 Equilibrium under Fee-for-Service and Balance Billing

In this section, the government purchaser sets the price p_g. In addition to receiving p_g from the purchaser for each service provided, GPs can charge an additional amount p_p that the patients pays. In this case, GPs receive $p = p_p + p_g$ and patients pay p_p. It is assumed that given p_g, p_p is determined by the forces of competitive supply and demand.[5]

4.1 Equilibrium Determination of p_p

In this subsection, equilibrium p_p and the number of patients treated by high quality GPs is determined for any given number of high quality GPs, G_H. N_H patients demand the services of high quality GPs if $\theta(N_H) \cdot (q_H - q_L) = p_p$. That is, if the N_H^{th} patient's valuation of the higher quality service is equal to the price he/she pays for it.

By assumption, high quality GPs operate in competitive markets and take the price $p = p_p + p_g$ as given. Therefore, they choose the number of services to supply by equating price to marginal cost, that is $p_p + p_g = c'_H(n_H)$. Total supply is $N_H = n_H \cdot G_H$ and so market supply is given by $p_p + p_g = c'_H(\frac{N_H}{G_H})$. Equating market demand and market supply and rearranging yields

$$\theta(N_H) \cdot (q_H - q_L) = c'_H(\frac{N_H}{G_H}) - p_g. \tag{9}$$

[5]In Australia, in 2010, 75% of GP attendances were bulk billed, Medicare Australia Statistics, 2010, Monthly and Quarterly Standard Reports. GPs tend to either bulk bill, $p_p = 0$, or balance bill, $p_p > 0$, all patients. In addition, bulk billing and balance billing GPs are often located in close geographic proximity. This suggests that balance billing GPs are able to charge $p_p > 0$ not because of monopoly power, but rather because the market for high quality GPs is competitive and the service they offer is different to that of low quality GPs.

Condition (9) is solved for $\tilde{N}_H$ and $\tilde{p}_p = \theta(\tilde{N}_H) \cdot (q_H - q_L)$. Patients with preference parameters in $[\theta(\tilde{N}_H), \overline{\theta}]$ choose the high quality service and pay a premium of $\tilde{p}_p$.

4.2 Determination of p_g

In equilibrium, patients are allocated between GPs so that (9) is satisfied. In addition, in equilibrium, GPs are allocated between low and high quality so that the utility of the marginal GP is equal whether or not she chooses low or high quality, that is,

$$p_g \cdot n_L - c_L(n_L) = (p_g + p_p) \cdot n_H - c_H(n_H) + \gamma(G_H) \cdot q_H. \qquad (10)$$

Using high quality GP market supply allows (10) to be rewritten as

$$p_g \cdot n_L - c_L(n_L) = c'_H(\frac{N_H}{G_H}) \cdot (\frac{N_H}{G_H}) - c_H(\frac{N_H}{G_H}) + \gamma(G_H) \cdot q_H. \qquad (11)$$

Given p_g and $N_L + N_H = \bar{N}$, (9) and (11) are solved simultaneously for equilibrium N_H and G_H.

Proposition 1: *If the government purchaser sets* $p_g = c'_L(\frac{\bar{N} - N^*_H}{\bar{G} - G^*_H})$, *then market determination of* p_p *yields the welfare maximising allocation of patients between GPs and the welfare maximising allocation of GPs between high and low quality.*

Proof: If $p_g = c'_L(\frac{\bar{N} - N^*_H}{\bar{G} - G^*_H})$, then profit maximisation yields $n_L = \frac{\bar{N} - N^*_H}{\bar{G} - G^*_H}$. Substituting this into (11) above yields

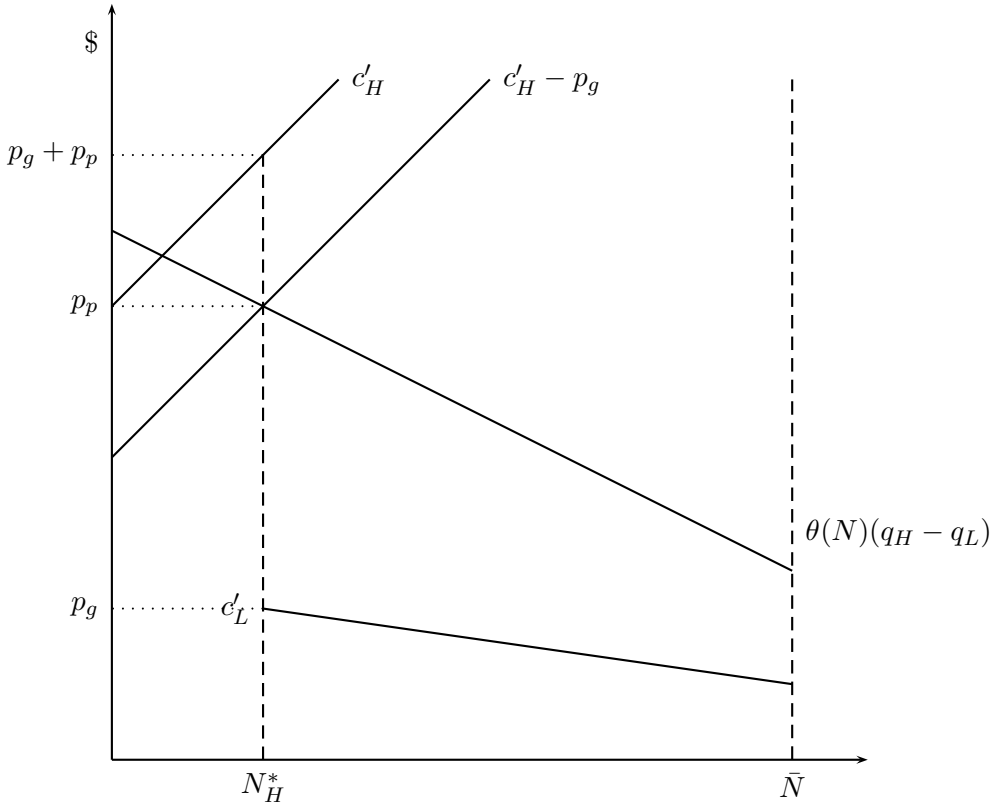
$$c'_L(\cdot) \cdot (\frac{\bar{N} - N^*_H}{\bar{G} - G^*_H}) - c_L(\frac{\bar{N} - N^*_H}{\bar{G} - G^*_H}) = c'_H(\cdot) \cdot (\frac{N_H}{G_H}) - c_H(\cdot) + \gamma(G_H) \cdot q_H. \quad (12)$$

Given the uniqueness of the solution to (7) and (8) and the fact that all patients are treated in the welfare maximum, the solution to (9) and (12) is N^*_H and G^*_H. □

Given G_H^*, the solution for N_H^* is shown in Figure 1. p_g is chosen so that $\bar{N} - N_H^*$ patients are treated by low quality GPs. The remaining patients pay $p_p = \theta(N_H^*)(q_H - q_L)$ and are treated by high quality GPs. By construction, condition (9) is satisfied at N_H^*.

Figure 1

Solution for N_H^*



The optimal government purchase price of $p_g = c'_L\left(\frac{\bar{N} - N_H^*}{G - G_H^*}\right)$ when combined with market determination of p_p duplicates the outcome that would arise if there were no government purchaser and the price paid to low quality GPs was market determined, the difference being that patients would pay

this price. Therefore, it is not surprising that balance billing leads to the welfare maximising (efficient) allocation of patients to GPs and the welfare maximisng (efficient) allocation of GPs between high and low quality.

5 Equilibrium under Fee-for-Service and No Balance Billing

In this section, it is assumed that patients make no payments, that is, $p_p = 0$. Let the price paid by the purchaser be denoted p_{nb}. Define p^0_{nb} as the highest price for which a high quality GP chooses $n_H = 0$. Assume that $\pi_L(p^0_{nb}) > \gamma(0)$ so that with p^0_{nb} all GPs choose to be low quality. That is, no GP chooses to be high quality unless they earn positive income. For $p_{nb} > p^0_{nb}$, as long as low quality GPs are not rationed, $\pi_L(p_{nb}) > \pi_H(p_{nb}) + \gamma(0)$ because $n_L > n_H$. In this case all GPs choose to be low quality.

Define p^1_{nb} by $\pi_L(p^1_{nb}) = \pi_H(p^1_{nb}) + \gamma(0)$, it is the lowest price at which the GP with the highest γ chooses high quality. For $p_{nb} \geq p^1_{nb}$, at least one GPs chooses high quality and low quality GPs are rationed. For $p_{nb} < p^1_{nb}$ all GPs choose low quality.

Let $s_H(p)$ be the supply function of a high quality GP and let $G_H(p)$ be a function that relates the number of high quality GPs to price. In addition, let the expected value of θ be $E\theta$. The number of patients that see a low quality GP is the residual $\frac{\bar{N} - s_H(p)\cdot(G_H)}{\bar{G} - G_H}$. Expected welfare is given by

$$\begin{aligned} W &= E\theta \cdot q_H \cdot s_H(p) \cdot G_H(p_{nb}) - c_H(s_H(p)) \cdot G_H(p_{nb}) \\ &+ E\theta \cdot q_L \cdot \left(\bar{N} - s_H(p) \cdot G_H(p_{nb})\right) - c_L\left(\frac{\bar{N} - s_H(p) \cdot G_H(p_{nb})}{\bar{G} - G_H(p_{nb})}\right) \cdot (\bar{G} - G_H(p_{nb})) \\ &+ \int_0^{G_H(p_{nb})} \gamma(G) \cdot q_H \cdot dG. \end{aligned} \tag{13}$$

The government purchaser chooses p_{nb} to maximise welfare. Let the solution

to this problem be given by p^*_{nb}. This discussion is summarised in the following propoition.

Proposition 2: *(i) If $p^*_{nb} < p^1_{nb}$, then all GPs choose low quality and the welfare maximum of Section 3 is not achieved. (ii) If $p^*_{nb} \geq p^1_{nb}$, then some GPs choose high quality and the welfare maximum of Section 3 is not achieved because patients are allocated randomly to high and low quality GPs.*

5.1 Single Price versus Balance Billing

Proposition 1 established that balance billing yields the efficient allocation of patients to GPs and the efficient allocation of GPs between high and low quality. This required two prices, one for high quality GPs, $p^*_g + p_p$, and one for low quality GPs, p^*_g. These two prices are essential in allocating patients to GPs since with one price patients end up being allocated randomly. This results in some patients who place a low value on a high quality service being treated by high quality GPs. In addition, with one price, GPs are allocated inefficiently between high and low quality as one instrument can not attain two objectives. The above discussion is summarised in the following proposition.

Proposition 3: Balance billing with $p_g = c'_L\big(\frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\big)$ and $p_p = \theta(N^*_H) \cdot (q_H - q_L)$ yields more welfare than no balance billing.

Once again this is not surprising. Essentially there are two competitive markets, the high and low quality market for GP services. Allowing two prices where the low quality price is optimally chosen and the high quality price is market determined is optimal as these two prices are able to allocate GPs between high and low quality and patients to GPs in an efficient manner. One price is unable to do this even if the allocation of GPs between high and low quality is fixed at the welfare maximum.

No Balance Billing: With no balance billing, if it is optimal to have some high quality GPs, then low quality GPs are rationed and have an incentive to over-service. This incentive is not present with balance billing as p_g is set at a level that induces low quality GPs to offer exactly the numbers of services demanded from them. This is a further advantage balance billing has over the optimal single price.

Balance Billing: With balance billing the choice of p_g by the purchaser is crucial. If $p_g < c'_L\big(\frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\big)$, then less patients are treated by both high and low quality GPs relative to the welfare maximum. Therefore, some patients, those who do not value quality highly, are not treated. For these patients, the benefit of treatment is greater than the marginal cost of being treated by a low quality GP and so low quality GPs can charge price p^L_p to these patients, where p^L_p is obtained from $p_g + p^L_p = c'_L\big(\frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\big)$. In this case, both high and low quality GPs balance bill with $p_p = \theta(N^*_H)\cdot(q_H - q_L) + p^L_p$ and welfare is maximised.

The analysis is more complicated if $p_g > c'_L\big(\frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\big)$. Totally differentiating (9) and (11), applying Cramer's Rule, and using the second order conditions for a welfare maximum yields

$$sign\left[\frac{\partial N_H}{\partial p_g}\right] = sign\left[c''_H(\cdot)\cdot\Big(\frac{N^*_H}{G^{*2}_H}\cdot\big(\frac{N^*_H}{G^*_H} - \frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\big)\Big) - \gamma'(G^*_H)\cdot q_H\right] \quad (14)$$

and

$$sign\left[\frac{\partial G_H}{\partial p_g}\right] = sign\left[\theta'(N^*_H)\cdot(q_H - q_L)\cdot\big(\frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\big) + \frac{c''_H(\cdot)}{G^*_H}\cdot\Big(\frac{N^*_H}{G^*_H} - \frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\Big)\right] \quad (15)$$

at N^*_H, G^*_H.

If $\left(\frac{N^*_H}{G^*_H} - \frac{\bar{N}-N^*_H}{\bar{G}-G^*_H}\right) < 0$, that is, at the welfare maximum, the number of patients treated by low quality GPs is greater than the number treated by high quality GPs, then $\frac{\partial G_H}{\partial p_g} < 0$ and there is an unambiguous excess supply

of low quality services at $p_g > c'_L\left(\frac{\bar{N}-N_H^*}{\bar{G}-G_H^*}\right)$. The intuition is clear, an increase in p_g increases the utility of being a low quality GP relative to being a high quality GP as low quality GPs provide more services than high quality GPs. Therefore, more GPs choose low quality. The number of patients is fixed, so with more low quality GPs and a higher price for their services there is an excess supply of low quality services and low quality GPs are rationed.

If $\left(\frac{N_H^*}{G_H^*} - \frac{\bar{N}-N_H^*}{\bar{G}-G_H^*}\right) > 0$, then $\frac{\partial N_H}{\partial p_g} > 0$ but the sign of $\frac{\partial G_H}{\partial p_g}$ is ambiguous. However, even if G_H increases, there is an excess supply of low quality services as the number of patients treated by low quality GPs is smaller than the number treated by high quality GPs and low quality GPs are rationed.

Therefore, if $p_g > c'_L\left(\frac{\bar{N}-N_H^*}{\bar{G}-G_H^*}\right)$ low quality GPs are rationed. This rationing may lead to over-servicing by low quality GPs as patients face a price of zero and low quality GPs want to provide more services than the rationed amount. The choice of p_g is therefore a crucial choice for the purchaser.[6]

So far, the interpretation of p_g is that it is the price a purchaser pays to GPs for providing a service. It provides insurance to patients and at the welfare maximum it provides complete insurance for those patients choosing low quality GPs and partial insurance for those patients choosing high quality. However, the model could be reinterpreted with patients paying p_g to obtain treatment from low quality GPs, or paying $p_g + p_p$ to obtain treatment from high quality GPs and then being reimbursed through insurance an amount p_i where $p_i = p_g$.

[6]Pauly (1991) argued that at the time US medicare prices for most services were greater than marginal cost and this resulted in an excess supply for most medicare services.

5.2 Local Markets

So far it has been assumed that there is only one market for GP services. It is now assumed that there are many local markets which are differentiated by having different distributions over patient preference parameters. This differentiation is achieved by adding a parameter α into the function that maps patients into θ, that is, $\theta(N;\alpha)$, where $\frac{\partial\theta}{\partial\alpha} \geq 0$. For a given N, the greater is α, the greater is θ. Assuming $\bar{N}$ is the same in all markets, the interpretation of one market having a higher α than another is that in the market with the higher α, the average patient has a stronger preference for high quality than in the market with the lower α.[7] Substituting $\theta(N;\alpha)$ for $\theta(N)$ in the definition of welfare, (3), results in $\theta(N_H)$ in first order condition (7) being replaced by $\theta(N_H;\alpha)$. Totally differentiating the new (7) and (8) and applying Cramer's rule yields

$$\begin{aligned} sign\left[\frac{\partial N_H}{\partial\alpha}\right] &= sign\left[-\frac{\partial\theta}{\partial\alpha}\cdot(q_H-ql)\cdot\left(-c''_H(\cdot)\cdot\frac{N_H^2}{G_H^3}\right.\right. \\ &- \left.\left. c''_L(\cdot)\frac{(\bar{N}-N_H)^2}{(\bar{G}-G_H)^3}+\gamma'(G_H)q_H\right)\right] > 0 \end{aligned} \tag{16}$$

and

$$sign\left[\frac{\partial G_H}{\partial\alpha}\right] = sign\left[\frac{\partial\theta}{\partial\alpha}\cdot(q_H-q_L)\cdot\left(c''_H(\cdot)\cdot\frac{N_H}{G_H^2}+c''_L(\cdot)\frac{(\bar{N}-N_H)}{(\bar{G}-G_H)^2}\right)\right] > 0 \tag{17}$$

The intuition is clear. In markets in which α is higher, patients have a greater preference for high quality and so the welfare maximum involves more patients being served by high quality GPs and more GPs of high quality than in markets in which α is lower.

[7] It could be that interval from which θ is drawn is unchanged and the cumulative distribution function associated with the higher α, $J(\theta)$, stochastically dominates $F(\theta)$, or it could be that the interval from which θ is drawn has an upper bound which is greater than $\overline{\theta}$.

Proposition 1 applies to all markets regardless of α, and so the welfare maximum in each market can be achieved by the purchaser setting $p_g = c'_L\big(\frac{\bar{N}-N^*_H(\alpha)}{\bar{G}-G^*_H(\alpha)}\big)$ and then letting the market determine p_p. By assumption, marginal cost is an increasing function and so $\frac{dp_g}{d\alpha}$ has the same sign as $\frac{d\frac{(\bar{N}-N^*_H(\alpha))}{(G-G^*_H(\alpha))}}{d\alpha}$. Now

$$\frac{d\frac{(\bar{N}-N^*_H(\alpha))}{(G-G^*_H(\alpha))}}{d\alpha} = \frac{-1}{\bar{G}-G^*_H} \cdot \frac{\partial N_H}{\partial \alpha} + \Big(\frac{\bar{N}-N_H(\alpha)}{(\bar{G}-G_H(\alpha))^2}\Big) \cdot \frac{\partial G_H}{\partial \alpha} \tag{18}$$

Substitution of (16) and (17) into (18) yields

$$sign\left[\frac{d\frac{(\bar{N}-N^*_H(\alpha))}{(G-G^*_H(\alpha))}}{d\alpha}\right] = sign\left[\gamma'(G_H)\cdot q_H - \Big(\frac{N_H}{G_H} - \frac{\bar{N}-N_H}{\bar{G}-G_H}\Big)\cdot \frac{N_H}{G_H^2}\cdot c''_H(\cdot)\right] \tag{19}$$

If $\frac{N_H}{G_H} > \frac{\bar{N}-N_H}{\bar{G}-G_H}$, then $\frac{d\frac{(\bar{N}-N^*_H(\alpha))}{(G-G^*_H(\alpha))}}{d\alpha} < 0$ and p_g is lower in the market with the greater α. On the other hand, if $\frac{N_H}{G_H} < \frac{\bar{N}-N_H}{\bar{G}-G_H}$, then p_g is higher in the market with the greater α for small $\gamma'(G_H)q_H$, but lower in the market with the greater α for large $\gamma'(G_H)q_H$.

Using similar analysis it can be shown that if $\frac{N_H}{G_H} < \frac{\bar{N}-N_H}{\bar{G}-G_H}$, then $\frac{d\frac{(N^*_H(\alpha))}{(G^*_H(\alpha))}}{d\alpha} > 0$ and $p_g + p_p$ is higher in the market with the higher α. These results are summarised in the following proposition.

Proposition 4: *In the welfare maximum, markets in which α is higher have more high quality GPs and more patients allocated to high quality GPs than markets in which α is lower. This welfare maximum can be implemented by the purchaser setting $p_g = c'_L\big(\frac{\bar{N}-N^*_H(\alpha)}{\bar{G}-G^*_H(\alpha)}\big)$ and then letting the market determine p_p. If $\frac{N_H}{G_H} < \frac{\bar{N}-N_H}{\bar{G}-G_H}$, then p_g is lower in the market with the greater α for large $\gamma'(G_H)q_H$. In this case, p_g+p_p is higher in the market with the greater α and so p_p is greater in the market with the greater α.*

The result that p_p is greater in the market where patients value high quality more greatly is intuitive, but depends on the conditions given in

Proposition 4. Therefore, although this market has more high quality GPs and more patients serviced by high quality GPs, the price patients pay for these high quality services may not be greater than in the market where high quality is valued less by patients.

Savage and Jones (2004) found that in Australia the bulk billing rate (the proportion of patients with $p_p = 0$) falls as local market average income increases. With the interpretation that local markets with higher average income have a higher average preference for high quality this result is consistent with Proposition 4, where an increase in α leads to an increase in N_H^*.

In general, to implement the welfare maximum in Proposition 4, the purchaser needs to set a different p_g in each market. However, in practice, purchasers usually set the same p_g in all markets independent of the distribution of preferences in those markets.[8] This means that in some local markets it will be set too low and in other markets it will be too high relative to the welfare maximum. As discussed in Section 5.1, this does not create a problem in markets in which it is set too low, but in markets in which it is set too high, low quality GPs have an incentive to over-service.

5.3 Two Conditions

The analysis of Section 4 is repeated for the case where patients can have one of two conditions, 1 or 2. The number of patients with condition 1 is $\bar{N}^1$ and with condition 2 is $\bar{N}^2$. Let q_i^k $k = 1, 2$; $i = H, L$ be the quality of service provided by a GP of quality i when the patient has condition k. From the patients perspective, if they have condition 1, it is assumed that seeing a high quality GP yields more utility than seeing a low quality GP,

[8] In Australia there is one bulk billing price p_g.

$q_H^1 > q_L^1$. However, if they have condition 2, it is assumed that they get the same utility from seeing a GP of high or low quality, $q_H^2 = q_L^2 = q^2$.[9]

Welfare Maximum: The welfare maximum involves all patients with condition 2 being serviced by low quality GPs as low quality GPs provide services at lower cost. In addition, the welfare maximum involves patients with condition 1 being allocated between GPs so that the following is satisfied,

$$\theta(N_H^1) \cdot (q_H^1 - q_L^1) = c'_H\big(\frac{N_H^1}{G_H^1}\big) - c'_L\big(\frac{\bar{N}^1 - N_H^1 + \bar{N}^2}{\bar{G} - G_H^1}\big). \tag{20}$$

This is very similar to (7) above. Patients with condition 1 are allocated between GPs so that the extra marginal benefit of allocating one more patient to a high quality GP, $\theta(N_H^1) \cdot (q_H^1 - q_L^1)$, is equal to the extra marginal cost of doing so, $c'_H\big(\frac{N_H^1}{G_H^1}\big) - c'_L\big(\frac{\bar{N}^1 - N_H^1 + \bar{N}^2}{\bar{G} - G_H^1}\big)$. Let the solution to (20) be given by N_H^{1*}. Patients with condition 1 and $\theta \geq \theta(N_H^{1*})$ see high quality GPs, patients with condition 1 and $\theta < \theta(N_H^{1*})$ and all patients with condition 2 see low quality GPs. With appropriate amendments a condition similar to (8) determines the welfare maximising allocation of GPs between high and low quality

Proposition 1 is easily extended to the case of two conditions in the following proposition.

Proposition 5: *If the government purchaser sets* $p_g = c'_L\big(\frac{\bar{N}^1 - N_H^1 + \bar{N}^2}{G - G_H^1}\big)$, *then market determination of* $p_p = \theta(N_H^{1*}) \cdot (q_H^1 - q_L^1)$ *yields the welfare maximising allocation of patients between GPs and the welfare maximising allocation of GPs between high and low quality.*

Balance billing with two conditions has interesting implications. With

[9] Condition 2 might be a condition the patient regularly suffers from and goes to a GP to get a prescription for a pharmaceutical. In this case, from the patients perspective, it does not matter whether a high or low quality GP writes the prescription.

the optimal p_g given in Proposition 5, a patient with $\theta \geq \theta(N_H^{1*})$ and condition 1 chooses to be treated by a high quality GP while the same individual with condition 2 chooses to be treated by a low quality GP. The welfare maximum has the same patient being treated by different quality GPs depending on the condition they have. Given patient histories are important in correctly diagnosing and treating patients and given the welfare maximum has patients being served by different GPs depending on condition, it is important that patient histories are available to all GPs.

6 Payment by Capitation

Under payment by capitation GPs receive a fixed payment, k_g, for each patient registered with them. This payment is made by the government purchaser. Once a patient is registered with a GP all primary healthcare services are provided by this GP. In addition, it is assumed that GPs can charge patients a price, k_p, per-service delivered. In this section, the case where $k_p > 0$ will be referred to as capitation coupled with balance billing.

It is assumed that patients register with GPs before they know whether they have a medical condition. The total number of patients is $\hat{N}$. A proportion ϕ of these are assumed to have the same medical condition. Therefore, there are $\bar{N} = \phi \cdot \hat{N}$ patients who need treatment.

Given $\bar{N}$, the welfare maximising allocation of GPs between high and low quality and of patients between high and low quality GPs is given in Section 3 above as the solution to (7) and (8).

6.1 Equilibrium Determination of k_p

Given G_H, $\hat{N}_H$ patients demand to be registered with high quality GPs if $\theta(\hat{N}_H) \cdot (q_H - q_L) = k_p$. High quality GPs choose the number of patients

to register, $\hat{n}_H$, to maximise income $y_H = k_g \cdot \hat{n}_H + k_p \cdot \phi \cdot \hat{n}_H - c_H(\phi \hat{n}_H)$, where $k_g \hat{n}_H$ is income from capitation and $k_p \cdot \phi \cdot \hat{n}_H - c_H(\phi \hat{n}_H)$ is income from providing services to $\phi \hat{n}_H$ sick patients. Income maximisation leads to the condition $k_g + \phi k_p = \phi \cdot c'_H(\phi \cdot \hat{n}_H)$. Since $\hat{n}_H = \frac{\hat{N}_H}{G_H}$, this condition can be written in terms of $\hat{N}_H$ as $k_g + \phi k_p = \phi \cdot c'_H(\frac{\phi \cdot \hat{N}_H}{G_H})$. Equating demand and supply of registrations yields

$$\theta(\hat{N}_H) \cdot (q_H - q_L) = c'_H(\frac{\phi \cdot \hat{N}_H}{G_H}) - \frac{k_g}{\phi}. \tag{21}$$

Condition (21) is solved for $\hat{N}_H^k$ and $k_p^k = \theta(\hat{N}_H^k) \cdot (q_H - q_L)$. In equilibrium, patients with preference parameters in $[\theta(\hat{N}_H^k), \overline{\theta}]$ register with high quality GPs and pay k_p^k for high quality GP services.

6.2 Determination of k_g

In equilibrium, the utility of the marginal GP is the same whether high or low quality is chosen, that is,

$$k_g \cdot \hat{n}_L - c_L(\phi \hat{n}_L) = k_g \cdot \hat{n}_H + k_p \cdot \phi \cdot \hat{n}_H - c_h(\phi \hat{n}_H) + \gamma(G_H) \cdot q_H \tag{22}$$

Using the income maximisation condition, this can be written as

$$k_g \cdot \hat{n}_L - c_L(\phi \hat{n}_L) = c'_H(\frac{\phi \hat{N}_H}{G_H}) \cdot (\frac{\phi \hat{N}_H}{G_H}) - c_h(\frac{\phi \hat{N}_H}{G_H}) + \gamma(G_H) \cdot q_H \tag{23}$$

Given, p_g and $\hat{N}_H + \hat{N}_L = \hat{N}$, (21) and (23) are solved for the equilibrium number of patients registered with high quality GPs, $\hat{N}_H$, and the equilibrium number of high quality GPs, $\hat{G}_H$.

In the following proposition it is assumed that $\theta(\hat{N}_H)$ drawn from the entire population of patients, $\hat{N}$, is the same as $\theta(\phi \hat{N}_H)$ drawn from the population of patients that actually have the condition, $\phi \hat{N}$.[10]

[10] Assumimg all patients that are registered with high quality GPs are equally likely to fall sick this is approximately true for large $\hat{N}_H$.

Proposition 6: *If the government purchaser sets* $k_g = \phi c'_L\left(\frac{\phi(\hat{N}-\hat{N}^*_H)}{\bar{G}-G^*_H}\right)$, *where* $\phi\hat{N}^*_H = N^*_H$, *then market determination of* k_p *yields the welfare maximising allocation of patients between GPs and the welfare maximising allocation of GPs between high and low quality.*

The proof is identical to that of Proposition 1.

The optimal government capitation payment $k_g = \phi c'_L\left(\frac{\phi(\hat{N}-\hat{N}^*_H)}{\bar{G}-G^*_H}\right)$ when coupled with market determination of the fee-for-service, k_p, duplicates the welfare maximum. This is not surprising because with patients only having one condition a capitation payment is like a fee-for-service and Proposition 1 established that fee-for-service with p_g optimally chosen duplicated the welfare maximum. Even if patients can have many conditions, as long as the difference in quality from seeing a high or low quality GPs is the same for each condition, then Proposition 6 applies with ϕ scaled up to reflect the proportion of the population with any condition.[11] So fee-for-service or a capitation payment, when coupled with balance billing, can achieve the welfare maximising allocation of GPs between high and low quality and the welfare maximising allocation of patients between GPs.

If patients have conditions for which the difference in quality from seeing a high or low quality GP is different, then Proposition 5 applies and the welfare maximum can be achieved by fee-for-service and balance billing. However, it is no longer the case that fee-for-service and capitation are equivalent. In the welfare maximum with two conditions, patients with relatively high $\theta's$ choose to be treated by a high quality GP for one condition and a low quality GP for the other. Under capitation, patients register with a GP and obtain all their treatment from that GP. They are not allowed

[11] ϕ maybe greater than 1. In this case, the number of conditions that require treatment $\bar{N}$ is greater than the number of patients, $\hat{N}$.

to seek treatment for different conditions from different GPs. Therefore, in terms of the allocation of patients to GPs, fee-for-service coupled with balance billing is superior to capitation coupled with balance billing.

In Section 5.1 above, it was argued that if p_g is set too high relative to its welfare maximising level, then low quality GPs are rationed and have an incentive to over-service. As the information required to set p_g optimally is difficult for the purchaser to obtain it is possible that p_g might be set too high. This is a problem for fee-for-service. However, under capitation low quality GPs do not have an incentive to over-service if the capitation payment is set too high relative to its welfare maximising level as GP income does not depend on the number of services delivered.

In summary, fee-for-service with p_g optimally chosen duplicates the welfare maximum even if patients have many conditions and the difference between high and low quality varies from condition to condition. This is not true for capitation payments as patients can not seek treatment from different GPs conditional on the condition they have. However, if p_g and k_g are set too high relative to the welfare maximum, then fee-for-service provides an incentive for low quality GPs to over-service while a capitation payment does not. Which payment scheme is best from the point of view of the purchaser depends on whether p_g is optimally chosen, if it is, then fee-for-service with balance billing is superior to capitation with balance billing.

7 Conclusion

In this paper, in an environment where GPs are of differing quality and heterogeneous patients have different preferences for quality, it is shown that fee-for-service coupled with balance billing or capitation payments coupled

with balance billing are superior GP payment schemes than fee-for-service or capitation payments alone. This is because both achieve the welfare maximising allocation of GPs between high and low quality and the welfare maximising allocation of patients between GPs. To some extent this is not surprising as without balance billing there are two objectives and only one instrument. The policy implication is clear, balance billing should be allowed as it promotes an efficient allocation of GPs and patients. Balance billing in not allowed in the United States. It is allowed in Australia, though recently GPs have been given incentives to increase the bulk billing rate and so reduce the prevalence of balance billing, Savage and Jones (2004).

Where patients have more than one condition it is shown that fee-for-service coupled with balance billing is superior to capitation payments even when coupled with balance billing as fee-for-service allows patients to seek treatment from GPs of differing quality conditional on the condition they have. This is not possible under capitation payments. However, a payment system that encourages patients to seek treatment from different GPs conditional on condition does require a system of centralized patient records to ensure treatment is consistent with patients' medical histories.

The model of the paper has been framed in terms of GPs of different quality. However, a natural reinterpretation of the model has nurse practitioners replacing low quality GPs and GPs providing the high quality service. In this setting, GPs balance bill and nurse practitioners do not. If a patient has a condition for which quality matters then they seek treatment from GPs, if not, then they seek treatment from nurse practitioners. Similarly, GPs could be viewed as low quality and specialists as high quality. Patients seek treatment from GPs for minor conditions and are not balance billed but seek treatment from specialist for serious conditions and are balanced

billed.

Finally it is shown that the purchaser choice of the fee is crucial in terms of the incentives it provides to over-servicing. If it is set too high, low quality GPs are rationed and have an incentive to over-service. This is not so with capitation payments. Therefore, the superiority of fee-for-service coupled with balance billing over capitation payments coupled with balance billing depends very much on the fee being chosen optimally.

8 References

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

Glazer, J. and T. G. McGuire, 1993, Should Physicians be Permitted to 'Balance Bill' Patients?, *Journal of Health Economics*, **11**, 239-258.

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

Karlsson, M., 2007, Quality Incentives for GPs in a Regulated Market, *Journal of Health Economics*, 26, 699-720.

Ma, C. A. and T. McGuire, 1997, Optimal health Insurance and Provider Payment, *American Economic Review*, **87**, 685-704.

Medicare Australia Statistics, 2010, Monthly and Quarterly Standard Reports, https://www.medicareaustralia.gov.au/statistics/mth_qtr_std_report.shtml, 2nd quarter 2010, Table 4, last accessed August 2010.

Mussa, M. and S. Rosen, 1978, Monopoly and product quality, *Journal of Economic Theory*, **18**, 301-317

Pauly, M. V., 1991, Fee Schedules and Utilization, in H. E. Frech III, editor, Regulating Doctors Fees: Competition, Benefits, and Control under Medicare, (AEI Press, Washington D.C.).

Savage, E. and G. Jones, 2004, An Analysis of the General Practice Access Scheme on GP Incomes, Bulk Billing and Consumer Copayments, *The Australian Economic Review*, **37**, 31-40.

Tirole, J., 1988, The Theory of Industrial Organization, MIT Press, Cambridge Mass.

9 Appendix

Second Order Conditions for Welfare Maximum:

$$\frac{\partial^2 W}{\partial N_H^2} = \theta'(N_H) \cdot (q_H - q_L) - \frac{c_H''(\cdot)}{G_H} - \frac{c_L''(\cdot)}{\bar{G} - G_H} < 0 \qquad \text{(A-1)}$$

$$\frac{\partial^2 W}{\partial G_H^2} = \gamma'(G_H) \cdot q_H - \frac{N_H^2}{G_H^3} \cdot c_H''(\cdot) - \frac{(\bar{N} - N_H)^2}{(\bar{G} - G_H)^3} \cdot c_L''(\cdot) < 0 \qquad \text{(A-2)}$$

$$\frac{\partial^2 W}{\partial N_H^2} \cdot \frac{\partial^2 W}{\partial G_H^2} - (\frac{\partial^2 W}{\partial G_H \partial N_H})^2 > 0 \qquad \text{(A-3)}$$

Conditions (A-1) and (A-2) hold because $\theta'(N_H) < 0$, $c_H'' > 0$, $c_L'' > 0$ and $\gamma'(G_H) < 0$. Condition (A-3) is assumed to hold.