

Incentive Implications of Budgets for General Practices

(preliminary version)

Luigi Siciliani* Hugh Gravelle†

13 November 2009

Abstract

We model the incentives which arise from introducing budgets for general practitioners (family doctors), a policy known in England as Practice Based Commissioning (PBC). We investigate the effect of PBC on the incentives of general practitioners and hospitals to refer and treat patients. We show that if hospitals are paid according to an activity-based funding mechanism of the DRG pricing type the introduction of PBC reduces the number of referrals. In contrast if hospitals are paid according to a fixed budget the introduction of PBC has no effect on the number of referrals. We also show that if the purchaser cannot pay directly for preventive effort, the optimal price paid by the general practitioner to the hospital is higher than the optimal price when preventive effort is also contractible.

Keywords: Referrals, GPs, Budgets, Hospitals.

JEL classification: I11, I18.

*Department of Economics and Related Studies; and Centre for Health Economics, University of York, Heslington, York YO10 5DD, UK; and C.E.P.R., 90-98 Goswell Street, London EC1V 7DB, UK. E-mail: ls24@york.ac.uk.

†National Primary Care Research and Development Centre, Centre for Health Economics, University of York, York YO10 5D, UK; E-mail: hg8@york.ac.uk.

1 Introduction

Over the past decades activity-based payments (of the DRG type) have been increasingly used in many health-care systems (Mossialos, 2002). Hospitals are paid on the basis of the volume and type of patients treated, rather than some form of fixed budgets. Policy makers often argue that such new method of payment is beneficial because it encourages providers to increase activity and to contain costs. On the other hand, it has also been pointed out that such new method might provide excessive incentives to admit patients. One recent policy in England, known as Practice Based Commissioning (PBC) has been introduced to counteract the strong incentives for hospitals to overtreat. The policy consists of providing general practitioners (family doctors) with budgets, and make them pay directly for hospital services.

In this study we describe a theoretical model which investigates the incentives which arise from introducing budgets for general practitioners. We investigate the effect of PBC on the incentives of general practitioners and hospitals to refer and treat patients. We show that if hospitals are paid according to an activity-based funding mechanism (of the DRG pricing type) the introduction of PBC reduces the number of referrals. In contrast if hospitals are paid according to a fixed budget the introduction of PBC has no effect on the number of referrals. We also show that if the purchaser cannot pay directly for preventive effort, the optimal price paid by the general practitioner to the hospital is higher than the optimal price when preventive effort is also contractible. This arises because a higher price both reduces referrals and encourages prevention.

We also show that the optimal price for the GP is generally different from the optimal price for the hospital, and typically lower. One, perhaps, surprising result is that the optimal price paid by the GP for each referral increases with the degree of altruism of the GP. This is in contrast to the literature on hospital incentives where higher altruism typically implies a lower price. In our model, higher altruism for the GP implies a higher incentive to refer, and therefore increases the scope for using the price to reduce such incentive.

The study contributes to the literature on optimal incentive schemes in the healthcare sector, and in particular the primary care sector. Dusheiko et al. (2006) investigate both theoretically and empirically the effect of fundholding practices (where general practitioners face a monetary price, and patients face a waiting-time price) on the demand of cataract hospital admissions when there is rationing by waiting. They find that fundholding practices have lower admission rates than non-fundholders ones. Barros and Martinez-Giralt (2003) investigate the effect of different payment schemes for family doctors and hospitals on preventive effort and number of referrals. Our main set up is similar to theirs but differs in the following dimensions. First, we assume that providers are, at least to some extent, altruistic. Second, it is hospitals rather than family doctors who have the final word on referral decisions, i.e. hospitals can send patients back to the family doctor, and the decisions are taken sequentially rather than simultaneously. Third, and most importantly, we focus on the effect of introducing practice-based commissioning, where the family doctor pays directly for hospital services, and on rewarding preventive effort explicitly (as under the Quality Outcome Framework in England).

Brekke, Nuscheler and Straume (2007) investigate the effect of introducing a gate-keeping system (where patients can visit a specialist only after being referred by a family doctor). They show that gatekeeping may generate excessive quality competition and too much specialization. Malcomson (2005) investigates the extent to which the introduction of a gatekeeper effectively reduces the number of specialists contacts so that specialist services are used appropriately.

Our main set up is such that patients with different severity are treated either by the GP or the hospital. Malcomson (2005), Siciliani (2006), Hafsteinsdottir and Siciliani (2009) also assume patients differing in severity but the choice is between two hospital treatments which differ in intensity (rather than patients having either hospital treatment or GP care). The study also contributes to the literature on optimal payment for hospitals (Shleifer, 1985; Ellis and McGuire, 1986; Dranove, 1987; Ma, 1994; Ellis, 1998; Chalkley and Malcomson, 1998a, 1998b; Rickman and McGuire, 1999; De Fraja, 2000; Jack, 2005),

which does not model explicitly the interaction with primary-care providers.

Section 2 sets out the model and derives the main results. Section 3 derives the first best and the optimal prices. Section 4 concludes.

2 The Model

We analyse the interaction between a representative general practitioner (GP) and a representative hospital. We assume that the GP is paid by capitation. The GP receives a payment t for every patient registered in the practice. We assume the practice list is exogenous and the GP has N patients registered in the practice. The revenues are then equal to tN .

With a probability $\delta(e)$ the patient is ill and goes to see the GP, and with probability $(1 - \delta(e))$ the patient is healthy. e is the amount of preventive effort that the GP exerts to reduce the probability of illness (ie checking blood pressure, giving advice on weight, high cholesterol, etc) and $\partial\delta(e)/\partial e := \delta_e(e) < 0$. Therefore, the number of patients who visits the GP in any given period is $N\delta(e)$. If patients are in good health they have a utility equal to G . If patients are sick, we assume that they vary in severity s , which is distributed with density function $f(s)$ and cumulative density function $F(s)$ over the support $[\underline{s}, \bar{s}]$.¹ The utility of a patient with severity s after receiving treatment is $b(s)$, which is defined more explicitly below. We assume $b(s) \leq G$, i.e. some patients recover fully (if $b(s) = G$) while others don't ($b(s) < G$).

Define z_1 and z_2 as two exogenous severity cut-off points such that $z_2 > z_1$ and $z_1, z_2 \in [\underline{s}, \bar{s}]$. There are three groups of patients: 1) patients with "low" severity $s \in [\underline{s}, z_1)$ can only receive GP treatment (patient suffer from minor health problem, for example flu) or the hospital cannot treat them; 2) patients with "high" severity $s \in (z_2, \bar{s}]$: these patients can only be treated by a hospital, i.e. there is no treatment the GP can offer them; 3) patients with "middle" severity $s \in [z_1, z_2]$: these patients can be treated both by the GP and the hospital. We assume $z_1 \geq \underline{s}$. As a special case $z_1 = \underline{s}$ (the hospital can always

¹To keep the presentation simple we assume that preventive effort affects only the probability of patients falling sick and not the distribution of severity $f(s)$.

treat patients with any severity). In contrast $z_2 < \bar{s}$: there is always a subset of patients that the GP cannot treat.² We assume that all the patients receive treatment either from the GP or the hospital, ie there are no patients who receive no treatment at all. For the hospital the set of possible patients is $s \in [z_1, \bar{s}]$ while for the GP it is $s \in [\underline{s}, z_2]$. The patients that can be treated by both the GP and the hospitals are patients with severity $s \in [z_1, z_2]$.

Define $c^h(s)$ as the cost for the hospital from treating a patient with severity s (where h is a reminder of hospital), and $c^{gp}(s)$ as the cost for the GP from treating a patient with severity s . We assume that cost increases with severity, ie $\partial c^h(s)/\partial s := c_s^h > 0$ and $\partial c^{gp}(s)/\partial s := c_s^{gp}(s) > 0$. Moreover, we assume that for patients with middle severity the cost of providing hospital care is higher than the cost of providing GP care, ie $c^h(s) > c^{gp}(s)$ for $s \in [z_1, z_2]$. We assume that $c^{gp}(\underline{s}) > 0$: we can interpret this as the minimum (for example time) cost for the GP when seeing the patient even if no treatment is required ("you have nothing: go home, and rest"). Figure 1 illustrates the cost functions.

[Figure 1 here]

Define $b^{gp}(s) > 0$ ($b^h(s) > 0$) as the benefit for a patient with severity s which receives care from the GP (hospital). We assume i) $b^h(s) \geq b^{gp}(s)$: hospital treatment is weakly more beneficial than GP treatment,³ and ii) $\partial [b^h(s) - b^{gp}(s)]/\partial s \geq 0$: the difference between hospital and GP treatment does not decrease with the severity of the patient.

The timing of the model is the following. At time 0, the GP chooses the preventive effort. At time 1 the GP either refers the patient to the hospital or treats the patient, and decides how much preventive effort to exert. At time 2 the hospital either treats the patient or sends the patient back to the GP. At time 3 the GP treats the patients who are sent back from the hospital. We solve by backward induction.

²We assume z_1 and z_2 to be exogenous. The choice of z_2 (the maximum severity the GP can treat) may be endogenous. For example the GP might invest in facilities that might enable her to treat more severe patients.

³We could alternative assume that $b^{gp}(\underline{s}) \geq b^h(\underline{s})$: for patients with lowest severity the GP treatment is more beneficial (hospital treatment wastes patients' time or is more aggressive, ie X-rays, but does not increase the benefit from treatment) and $b^h(\bar{s}) \geq b^{gp}(\bar{s})$: for patients with highest severity the hospital treatment is more beneficial. This set-up would not qualitatively alter the main results of the model.

Stage 3 is trivial. If some patients are sent back to the GP, then the GP has to treat them. In the next section we investigate the incentive for the hospital to treat patients.

2.1 The Hospital (stage 2)

We investigate hospital's incentives to treat patients under two regimes. First, we assume that the hospital is paid according to an activity-based financing mechanism of the DRG type (in England this is often referred to as "Payment by Results"), where the hospital receives a price for every patient treated.⁴ The second (older) financing mechanism is where hospitals are paid by fixed budgets. Under both scenarios we assume that the hospital has discretion over which patients to treat and which patients to send back to the GP.

2.1.1 Activity-based financing

The hospital receives a price p^h for every patient treated. We assume that hospitals cannot refuse treatment to patients whom the GP cannot treat, i.e. patients with severity above z_2 . However, the hospital can (potentially) send back to the GP patients for whom the GP can offer a treatment, i.e. patients with severity weakly below z_2 .

Define z^{gp} as the *referral threshold of the GP*, i.e. the severity cut-off point such that all the patients whose severity is above z^{gp} are referred to the hospital. The total number of referrals is then equal to $R(z^{gp}) = [1 - F(z^{gp})] N\delta(e)$.

Does the hospital have an incentive to treat all the patients which are referred? Define z^h as the *treatment threshold of the hospital*, i.e. the severity cut-off point of the hospital above which patients are treated, and below which they are sent back to the GP.

The profit of the hospital is then equal to:

$$\pi^h(z^h) = N\delta(e) \left[p^h [1 - F(z^h)] - \int_{z^h}^{\bar{s}} c^h(s) f(s) ds \right] \quad (1)$$

⁴In England HRGs (Healthcare Resource Groups) are used for hospital payments, rather than DRGs (Diagnosis Related Groups).

We assume that hospitals are to some extent altruistic. In line with Chalkley and Malcolmson (1998) we capture altruism with the parameter $\alpha^h \in [0, 1]$. The utility function of the hospital is given by: $U^h(z^h) = \alpha^h B^h(z^h) + \gamma^h \pi^h(z^h)$, where:

$$B^h(z^h) = N\delta(e) \int_{z^h}^{\bar{s}} b^h(s) f(s) ds + N[1 - \delta(e)]G. \quad (2)$$

Therefore, we assume that the hospital cares only about the patients treated in the hospital, and not all the patients (i.e. also the those treated by the GP).⁵ The parameter $\gamma^h \in (0, 1)$ captures the degree of profits appropriability. With for-profit hospitals $\gamma^h = 1$. For non-profit institutions or public hospitals there may be constraints to the use of profits. Even if profits cannot formally be distributed, they can still be utilised to increase perks for hospital staff.

In the following we assume that $p^h \geq c^h(z_2)$, i.e. the price received by the hospital is higher than the cost of treating patients with severity z_2 (the lowest severity of patients who can receive treatment only in hospitals). We further discuss the role of this assumption below.

What is the effect of an increase in the number of patients treated on hospital utility (i.e. a reduction in z^h)? Differentiating, we obtain:

$$\frac{\partial U^h(z^h)}{\partial z^h} = - \left\{ \alpha^h b^h(z^h) + \gamma^h [p^h - c^h(z^h)] \right\} f(z^h) N \delta(e) < 0. \quad (3)$$

An increase in patients treated always increases the utility of the hospital. A marginal increase in the number of patients treated increases the altruistic benefit component, and also profits since the patients who are referred at the margin are those with lower severity and therefore lower costs. We therefore conclude that the hospital treats all the patients who are referred, i.e. $z^{h*} = z^{gp}$, where the notation $(.)^*$ is used to denote the optimal cut-off point (note that $c^h(z_2) \geq c^h(z^{gp})$ and therefore $p^h > c^h(z^{gp})$ for any $z^{gp} < z_2$). In

⁵We may alternatively assume that the hospital cares about all patients, regardless of who provides the treatment. The main results of the model would be qualitatively unaffected.

summary, there is no conflict in this case between the GP and the hospital. The hospital has no incentive to send patients back to the GP.

Above, we have assumed that $p^h \geq c^h(z_2)$. This is a sufficient condition. It guarantees that even if the hospital is profit maximiser, ie $\alpha^h = 0$, then $\partial U^h(z^h) / \partial z^h = -\gamma^h [p^h - c^h(z^h)] f(z^h) < 0$. Otherwise (if $p^h < c^h(z_2)$), the hospital would have no incentive to treat patients whose severity is between $s \in (\tilde{s}, z_2)$ where \tilde{s} is such that $p^h = c^h(\tilde{s})$, but only patients whose severity is $s \in [z_2, \bar{s}]$ and $s \in [z^h, \tilde{s}]$. Note that even in this case we have that for $z^h < \tilde{s}$, we have $\partial U^h(z^h) / \partial z^h < 0$. In words, if the hospital is profit maximiser and the tariff p^h is sufficiently low the hospital will: i) treat all the patients with high severity (whose severity is above z_2); ii) treat all patients with low severity (whose severity is below \tilde{s}), and, iii) send patients with middle severity back to the GP (whose severity is between \tilde{s} and z_2).

If the provider is to some extent altruistic, all patients with severity $s \in [z^{gp}, \bar{s}]$ will still receive treatment even if $p^h < c^h(z_2)$ as long as: $\alpha^h b^h(z^h) > \gamma^h [c^h(z^h) - p^h]$, i.e. the altruistic component is sufficiently high compared to the negative profit margin.

2.1.2 Fixed budget

We now assume that the hospital is remunerated according to a fixed budget T^h . We have shown above that under activity-based financing the hospital has an incentive to treat all the patients who are referred from the GP. This is not the case anymore under a fixed-budget rule. Suppose, as before, that the GP refers $R(z^{gp}) = (1 - F(z^{gp}))N\delta(e)$ patients, where z^{gp} is the severity cut-off point above which patients are referred to the hospital. Has the hospital incentive to treat all referred patients, as under activity-based financing?

The utility of the hospital is in this case:

$$U^h(z^h) = \alpha^h N\delta(e) \int_{z^h}^{\bar{s}} b^h(s) f(s) ds + \gamma^h \left[T^h - N\delta(e) \int_{z^h}^{\bar{s}} c^h(s) f(s) ds \right] \quad (4)$$

Differentiating, we obtain:

$$\frac{\partial U^h(z^h)}{\partial z^h} = \left[\gamma^h c^h(z^h) - \alpha^h b^h(z^h) \right] f(z^h) N \delta(e). \quad (5)$$

Notice that for altruism equal to zero, it is always the case that $\frac{\partial U^h(z^h)}{\partial z^h} > 0$. This is still the case if altruism is positive but sufficiently small. In this and the following paragraph, we assume that this is the case (we discuss below the case of high altruism). It follows that the cut-off point is the highest possible compatible with the constraint $z^h \leq z_2$, ie $z^{h*} = z_2$. The hospital treats only the patients whose severity is above z_2 (ie those who can be treated only by the hospital).

Suppose that $z^{gp} < z^h = z_2$. In this case, the hospital and the GP have a disagreement. The GP would like to refer all patients whose severity is above z^{gp} and treat only those below. The hospital would like to treat only those patients whose severity is above z_2 . We assume that the hospital has all the bargaining power in terms of treatment decisions. All the patients who are not treated are sent back to the GP. This assumption seems reasonable as GPs have no power to oblige hospitals to treat patients. Ultimately, it is the hospital who has to provide the treatment.⁶

In summary, if altruism is sufficiently small, then for a given referral cut-off point $z^{gp} < z^{h*} = z_2$. The hospital treats all patients whose severity is above z_2 , while patients with lower severity with $s \in [z^{gp}, z_2]$ are sent back to the GP.

Suppose now that altruism is high. Then an interior solution may exist and the optimal cut-off point for the hospital is such that $\frac{\partial U^h(z^h)}{\partial z^h} = 0$, or, more extensively,

$$\gamma^h c^h(z^{h*}) f(z^{h*}) N \delta(e) = \alpha^h b^h(z^{h*}) f(z^{h*}) N \delta(e). \quad (6)$$

The optimal cut-off point for the hospital is such that the marginal cost of treatment is equal to the marginal benefit weighted by the altruistic component.⁷ Notice that if

⁶The case where the GP has positive but small bargaining power could be modelled as Nash bargaining. In this case, the hospital would to treat some extra patients with a cut-off point which is intermediate between z^{gp} and $z^h = z_2$.

⁷The Second Order Condition is: $\gamma^h c_s^h(z^h) f(z^h) N \delta(e) - \alpha^h b_s^h(z^h) f(z^h) N \delta(e) < 0$, which is always

$z^h > z_2$ then the constraint $z^h \leq z_2$ will be binding and $z^{h*} = z_2$: the analysis is the same as for low altruism. Otherwise, $z^{h*} < z_2$. In this case the hospital is willing to treat more than $R(z_2) = (1 - F(z_2))N\delta(e)$ patients. If z^{gp} is sufficiently low (ie below z^{h*}), then patients whose severity is above z^{h*} receive hospital treatment while patients with severity $s \in [z^{gp}, z^{h*}]$ are sent back to the GP. In contrast, if z^{gp} is sufficiently high (ie above z^{h*}), then the hospital is willing to treat more patients than the ones who are referred. In this case the hospital treats only patients with severity above z^{gp} .

2.2 The general practitioner (stage 1)

We investigate the optimal referral decisions for the representative GP. Define p^{gp} as the price that each GP has to pay for each referral to the hospital. In the absence of Practice Based Commissioning the price $p^{gp} = 0$ and the GP pays nothing for a referral. In contrast, under Practice Based Commissioning, the GP has to pay a price $p^{gp} > 0$ to the hospital for each patient referred and treated by the hospital. In this case, the GP also receives an additional budget T^{gp} on top of the capitation payment t . We also assume that the GP has to incur a (possibly small) administrative cost k for each referral.

Potentially, the price paid by the GP p^{gp} may differ from the price received by the hospital. We therefore present the analysis for the general case $p^h \neq p^{gp}$, and then discuss the (realistic) special case where $p^h = p^{gp}$, ie the price paid by the GP is equal to the price received by the hospital (which is in line with what is currently observed in England). Note that if $p^h > p^{gp}$ then the purchaser (or Department of Health) will fund the difference between the two tariffs.

We also consider the possibility that preventive effort exerted by the GP is contractible. This is in line with the QOF (Quality and Outcomes Framework) initiative recently implemented in England, which rewards higher performance on several observable indicators. Analytically, we assume that a price p^e is paid for each unit of effort exerted. Below, we discuss the special case where $p^e = 0$, which corresponds to countries where such payments systems are not in place.

satisfied when the cost is sufficiently flat compared to the benefit function.

As in the previous section we distinguish between the case where the hospital is paid according to activity-based financing and to a fixed budget.

2.2.1 Activity-based financing

In this case, the GP anticipates that under activity-based funding the hospital is willing to treat all the patients who are referred. Therefore, the GP can simply choose the severity threshold which maximises her utility function.

The utility of the GP is equal to:

$$\begin{aligned}
 U^{gp}(z^{gp}, e) = & tN + \alpha^{gp} B^{gp}(z^{gp}) + p^e e - N\delta(e) \int_{\underline{s}}^{z^{gp}} c^{gp}(s) f(s) ds - \phi(e) \quad (7) \\
 & + \gamma^{gp} \left[T^{gp} - p^{gp} N\delta(e) \int_{z^{gp}}^{\bar{s}} f(s) ds \right] - kN\delta(e) \int_{z^{gp}}^{\bar{s}} f(s) ds
 \end{aligned}$$

where z^{gp} is the cut-off point over which the patient is referred to the hospital, and e is preventive effort. z^{gp} and e are the choice variables of the GP. α^{gp} and γ^{gp} are positive parameters respectively indicating the degree of altruism and the degree of appropriability of profits of the GP, and

$$B^{gp}(z^{gp}, e) = N\delta(e) \left[\int_{\underline{s}}^{z^{gp}} b^{gp}(s) f(s) ds + \int_{z^{gp}}^{\bar{s}} b^h(s) f(s) ds \right] + N [1 - \delta(e)] G. \quad (8)$$

We therefore assume that the GP cares about the benefit of all her patients regardless of who provides the treatment (the GP herself or the hospital). If there is no Practice Based Commissioning, then $T^{gp} = p^{gp} = 0$. Under PBC the GP receives an additional budget T^{gp} but has to pay the price p^{gp} for each referral.

The optimal effort and cut-off point, if an interior solution exists, is such that $\frac{\partial U^{gp}}{\partial e} = 0$,

$\frac{\partial U^{gp}}{\partial z^{gp}} = 0$, or more extensively:

$$p^e - N\delta_e(e^*) \left[\begin{array}{c} \int_{\underline{s}}^{z^{gp*}} c^{gp}(s)f(s)ds + (\gamma^{gp}p^{gp} + k) \int_{z^{gp*}}^{\bar{s}} f(s)ds \\ +\alpha^{gp} \left(G - \int_{\underline{s}}^{z^{gp*}} b^{gp}(s)f(s)ds - \int_{z^{gp*}}^{\bar{s}} b^h(s)f(s)ds \right) \end{array} \right] = \phi_e(e^*) \quad (9)$$

$$\gamma^{gp}(p^{gp} + k)f(z^{gp*})N\delta(e^*) = \left\{ \alpha^{gp} [b^h(z^{gp*}) - b^{gp}(z^{gp*})] + \gamma^{gp}c^{gp}(z^{gp*}) \right\} f(z^{gp*})N\delta(e^*) \quad (10)$$

The optimal level of preventive effort is such that the marginal benefit is equal to the marginal disutility from effort. The marginal benefit has three components: i) the price p^e (like in the QOF in England); ii) the health benefit for the individuals registered in the practice from a reduced probability of falling ill, weighted by the altruistic component; iii) the savings for the GP practice from reduced number of referrals, which implies lower direct cost of provision (and administrative costs) and reduced payments to the hospital.

If an interior solution exists, the optimal cut-off point is such that the marginal benefit from saving the price p^{gp} is equal to the marginal cost of treatment for the GP plus the net marginal benefit from hospital treatment weighted by altruism.⁸ Note that the optimal cut-off point is independent of the optimal preventive effort, and viceversa ($\partial^2 U^{gp}(z^{gp*}, e^*) / \partial z^{gp} \partial e = \partial^2 U^{gp}(z^{gp*}, e^*) / \partial e \partial z^{gp} = 0$).

It is useful to start by investigating the special case where altruism is zero, and there is no PBC, i.e. $p^{gp} = 0$. In this case, a corner solution arises as $\partial U^{gp}(z^{gp}) / \partial z^{gp} = -\gamma^{gp}c^{gp}(z^{gp})f(z^{gp}) < 0$. The GP has an incentive to reduce the cut-off point as much as possible, ie to refer all patients whose severity is above z_1 , ie with $s \in [z_1, \bar{s}]$. In summary,

⁸The SOCs are: $\partial^2 U^{gp}(z^{gp}) / \partial^2 z^{gp} = -N\delta(e) \{ \alpha^{gp} [b_s^h(z^h) - b_s^{gp}(z^h)] + \gamma^{gp}c_s^{gp}(z^{gp*}) \} f(z^{gp*}) < 0$;
 $\partial^2 U^{gp}(e) / \partial^2 e = -N\delta_{ee}(e^*) \left[\begin{array}{c} \int_{\underline{s}}^{z^{gp*}} c^{gp}(s)f(s)ds + (p^{gp} + k) \int_{z^{gp*}}^{\bar{s}} f(s)ds \\ +\alpha^{gp} \left(G - \int_{\underline{s}}^{z^{gp*}} b^{gp}(s)f(s)ds - \int_{z^{gp*}}^{\bar{s}} b^h(s)f(s)ds \right) \end{array} \right] - \phi_{ee}(e^*) < 0$;
 $\partial^2 U^{gp}(z^{gp*}, e^*) / \partial z^{gp} \partial e = \partial^2 U^{gp}(z^{gp*}, e^*) / \partial e \partial z^{gp} = 0$.

since the hospital treats all the patients who are referred, then patients with severity $s \in [z_1, \bar{s}]$ receive treatment, while patients with severity $s \in [s, z_1)$ are treated by the GP. Notice that adding altruism does not change the result. Since hospital treatment is more beneficial, the GP has an extra motive to refer the patient to the hospital in the absence of financial incentives. Analytically,

$$\partial U^{gp} (z^{gp}) / \partial z^{gp} = - \left\{ \gamma^{gp} c^{gp}(z^{gp}) + \alpha^{gp} \left[b^h(z^{gp}) - b^{gp}(z^{gp}) \right] \right\} f(z^{gp}) < 0$$

What is the effect of introducing PBC? If the price p^{gp} is positive and sufficiently high the cut-off point $z^{gp*} > z_1$. Patients with severity $s \in [z^{gp*}, \bar{s}]$ receive treatment, while patients with severity $s \in [s, z^{gp*})$ are treated by the GP. (Note that if the price is sufficiently small, a corner solution still arises, ie the introduction of a small positive price has no effect on the number of referrals). Moreover, we have: $\partial z^{gp*} / \partial p^{gp} = \partial^2 U^{gp} / \partial z^{gp} \partial p^{gp} / (-\partial^2 U^{gp} / \partial (z^{gp})^2) > 0$. The total number of hospital referrals is equal to $R(p^{gp}) = \int_{z^{gp*}}^{\bar{s}} f(s) ds = 1 - F(z^{gp*})$, with $\frac{\partial R}{\partial p^{gp}} = -f(z^{gp*}) \frac{\partial z^{gp*}}{\partial p^{gp}} < 0$. A higher price p^{gp} therefore reduces the number of referrals. An increase in price p^{gp} also tends to increase preventive effort, as the GP saves more money for a marginal increase in preventive effort. Analytically, $\partial e^* / \partial p^{gp} = \partial^2 U^{gp} / \partial e \partial p^{gp} / (-\partial^2 U^{gp} / \partial e^2) > 0$.

The following proposition summarises the main results of this section.

Proposition 1 *If the hospital is paid according to an activity-based financing mechanism, the introduction of practice based commissioning reduces the number of hospital referrals and treatments if the price that the GP has to pay is sufficiently high. Moreover, it increases preventive effort.*

A corollary of the proposition is that in the absence of PBC, the number of referrals and treatments is equal to $N\delta(e)(1 - F(z_1))$: all patients that could potentially be treated by the GP are treated by the hospital.

Finally, note that a higher price for preventive effort encourages higher effort, as intuitively expected, and has no effect on the optimal cut-off point.

2.2.2 Fixed budgets

Suppose now that the hospital is paid according to a fixed budget and that there is no PBC (i.e. $p^{gp} = 0$) and no QOF (i.e. $p^e = 0$). In this case the GP still would like to refer all the patients with severity above z_1 . However, the GP anticipates that all the patients with severity below z_2 will be sent back to the GP. Since making a hospital referral has a cost k (for example filing a referral report), the GP does not have an incentive to make referrals for patients who will later be sent back to the GP without treatment. Therefore, the GP refers to the hospital only patients whose severity is above z_2 .

We have so far assumed that the price paid by the GP is zero (ie $p^{gp} = 0$). Introducing a positive price would not alter the conclusion, since the GP is already referring the minimum possible number of patients to the hospital. In this case PBC would have no effect on the number of referrals. This prediction is in line with the argument that PBC was introduced to counter the "excessive" incentives for hospitals to treat patients induced by an activity-based funding system for the hospital (of the DRG type).

As in the previous section, an increase in the price p^e increases preventive effort.

3 First Best

We assume that the purchaser of health services is utilitarian and maximises the sum of patients's benefit and the utility of the providers (the GP and the hospital) net of the transfers to the providers weighted by the opportunity cost of public funds. To avoid double-counting we exclude the altruistic component in the utility function of the provider (Hammond, 1987; Chalkey and Malcomson, 1998b).

We assume that the optimal design of the payment system has to satisfy the participation constraint of each provider (the GP and the hospital), as well as a limited-liability constraint. In words, we assume that not only providers need to have a weakly positive utility, but also a weakly positive profit (i.e. providers cannot go bankrupt). It is straightforward to show that whenever the limited-liability constraint is satisfied, the participation constraint is also satisfied.

Define λ as the opportunity cost of public funds. The utilitarian welfare function for a given severity cut-off point is equal to:

$$W(z) = N\delta(e) \left[\int_{z_1}^z b^{gp}(s)f(s)ds + \int_z^{\bar{s}} b^h(s)f(s)ds \right] + \pi^{gp}(z) + \pi^h(z) \quad (11)$$

$$-(1 + \lambda) \left[tN + p^e e + T^{gp} + T^h + (p^h - p^{gp})N\delta(e)(1 - F(z)) \right]$$

where

$$\pi^h(z) = T^h + N\delta(e) \left[p^h (1 - F(z)) - \int_z^{\bar{s}} c^h(s)f(s)ds \right]$$

$$\pi^{gp}(z) = tN + p^e e + T^{gp} - (p^{gp} + k)N\delta(e)(1 - F(z)) - N\delta(e) \int_{\underline{s}}^{z^{gp}} c^{gp}(s)f(s)ds - \phi(e)$$

After substitution, we can re-write the welfare function as:

$$W = N\delta(e) \left[\int_{z_1}^z b^{gp}(s)f(s) + \int_z^{\bar{s}} b^h(s)f(s) \right] - \lambda\pi^{gp}(z) - \lambda\pi^h(z) \quad (12)$$

$$-(1 + \lambda) \left\{ N\delta(e) \left[\int_z^{\bar{s}} [c^h(s) + k]f(s)ds + \int_{\underline{s}}^{z^{gp}} c^{gp}(s)f(s)ds \right] - \phi(e) \right\}$$

Since leaving a positive profit is always costly for welfare, it is optimal to set profits to zero.⁹ The purchaser's problem is to set the cut-off point and preventive effort such that it maximises the difference between the benefit for the patients and the costs for the provider weighted by the opportunity cost of public funds. The optimal level of prevention and the optimal cut-off point are such that:

$$-\delta_e(\hat{e})N \left[G - \left(\int_{\underline{s}}^{\hat{z}} b^{gp}(s)f(s)ds + \int_{\hat{z}}^{\bar{s}} b^h(s)f(s)ds \right) \right] = (1 + \lambda)\phi_e(\hat{e}) \quad (13)$$

⁹As we show below, the purchaser has enough instruments to avoid leaving rents (the capitation fee t ; prices p^{gp} and p^h or budgets T^h and T^{gp}).

$$N\delta(\hat{e})f(\hat{z}) \left\{ b^{gp}(\hat{z}) - b^h(\hat{z}) - (1 + \lambda) \left[c^{gp}(\hat{z}) - (c^h(\hat{z}) + k) \right] \right\} = 0 \quad (14)$$

The optimal amount of preventive effort is such that the marginal benefit for the patient from a reduced probability of illness is equal to the marginal disutility for the GP. The optimal cut-off point can be re-written more intuitively as:

$$b^h(\hat{z}) - b^{gp}(\hat{z}) = (1 + \lambda) \left[(c^h(\hat{z}) + k) - c^{gp}(\hat{z}) \right] \quad (15)$$

The optimal cut-off point is such that the extra benefit from hospital care is equal to the additional cost from hospital care (as opposed to GP care). Notice that for the SOC to be satisfied we need: $N\delta(e) [b_s^{gp}(z) - b_s^h(z)] - (1 + \lambda) N\delta(e) [c_s^{gp}(z) - c_s^h(z)] < 0$. This is certainly satisfied if $c_s^{gp}(s) > c_s^h(s)$, ie the cost of the GP increases faster with severity than for the hospital, which seems plausible, and if $[b_s^{gp}(z) - b_s^h(z)] < 0$, or $b_s^h(z) - b_s^{gp}(z) > 0$ ie the difference in benefit between hospital and GP care increases with severity, which also seems plausible.

Figure 2 illustrates the optimal solution.

[Figure 2 here]

3.1 Implementation

Suppose that the hospital is paid according to activity-based financing. The first best can be obtained by setting the prices p^{gp} , p^e and p^h in the following way. The optimal price for a referral is equal to:

$$\hat{p}^{gp} = c^h(\hat{z}) + [b^h(\hat{z}) - b^{gp}(\hat{z})] \left(\frac{\alpha^{gp}}{\gamma^{gp}} - \frac{1}{1 + \lambda} \right) \quad (16)$$

The optimal referral price for the GP is equal to the marginal cost of treatment for the hospital, plus the difference in the marginal benefit between hospital and GP treatment weighted by the degree of altruism, the opportunity cost of public funds, and the degree of appropriability of profits.

If altruism is equal to zero, then: $\hat{p}^{gp} = c^h(\hat{z}) - [b^h(\hat{z}) - b^{gp}(\hat{z})] \frac{1}{1+\lambda} < c^h(\hat{z})$ so that the price is below the marginal cost of treatment for the hospital. It is only for high altruism that the price is above the marginal cost. An interesting special case is when altruism is one, the opportunity cost of public funds is zero, and the degree of appropriability of profits is equal to one, in which case the referral price for the GP is exactly identical to the marginal cost of the hospital, $\hat{p}^{gp} = c^h(\hat{z})$. Perhaps counterintuitively, higher altruism generally reduces the optimal price. Since higher altruism increases the incentives for the GP to refer patients, the price for referrals is raised to counter such effect.

The fact that the optimal price of a referral is typically positive also implies that if the hospital is paid according to activity-based financing with no Practice Based Commissioning, then the number of referrals and treatments is excessively high compared to the first best. However, if PBC is introduced, and since $\frac{\partial z^{gp*}}{\partial p^{gp}} > 0$, there exists a price \hat{p}^{gp} which implements the first best.

Suppose that $T^h = 0$, i.e. the hospital receives no fixed budget component. The optimal payment for the hospital is such that:

$$\hat{p}^h = \frac{\int_{\hat{z}}^{\bar{s}} c^h(s) f(s) ds}{1 - F(\hat{z})} > c^h(\hat{z}) \quad (17)$$

The optimal price for the hospital is equal to the average treatment cost, so that the profit is equal to zero. Since the marginal patient is the one with lowest severity, the average price is above the marginal cost. Note also that the optimal price for the hospital is in general different from the price of referral. The price for the hospital is higher whenever:

$$\frac{\int_{\hat{z}}^{\bar{s}} c^h(s) f(s) ds}{1 - F(\hat{z})} - c^h(\hat{z}) > [b^h(\hat{z}) - b^{gp}(\hat{z})] \left(\frac{\alpha^{gp}}{\gamma^{gp}} - \frac{1}{1+\lambda} \right) \quad (18)$$

This is clearly the case for sufficiently low altruism.

Suppose now that $T^h \neq 0$, i.e. the purchaser can make use of a fixed-budget com-

ponent. Then, the first best can also be obtained by setting $\widehat{p}^h = \widehat{p}^{gp} = c^h(\widehat{z}) - [b^h(\widehat{z}) - b^{gp}(\widehat{z})] \frac{1}{1+\lambda} \left(\frac{\alpha^{gp}}{\gamma^{gp}} - \frac{1}{1+\lambda} \right)$ and setting:

$$T^h = \int_{\widehat{z}}^{\bar{s}} [c^h(s) - \widehat{p}^{gp}] f(s) ds > 0. \quad (19)$$

This type of payment system for hospitals is for example observed in Norway, where hospitals are paid a tariff which corresponds to around 40-60% of the average cost, and the remaining according to a fixed budget.

The optimal price for prevention is:

$$\begin{aligned} \widehat{p}^e = & (1 - \alpha^{gp}) \left[G - \left(\int_{\underline{s}}^{\widehat{z}} b^{gp}(s) f(s) ds + \int_{\widehat{z}}^{\bar{s}} b^h(s) f(s) ds \right) \right] \\ & - \left[\int_{\underline{s}}^{\widehat{z}} c^{gp}(s) f(s) ds + (\gamma^{gp} \widehat{p}^{gp} + k) \int_{\widehat{z}}^{\bar{s}} f(s) ds \right] - \lambda \phi_e(\widehat{e}) \end{aligned} \quad (20)$$

The optimal price is positive for sufficiently low altruism. Note that even if altruism is equal to zero, the optimal price is below the marginal patient's benefit from prevention (first term). This arises because the GP can reduce costs by increasing prevention effort (second term) and also because the GP does not take into account the additional cost due to the opportunity cost of public funds (third term).

For sufficiently high altruism, the optimal price is negative. If prices cannot be negative, which seems plausible, even a price for prevention equal to zero will imply an excessively high level of prevention. A positive price is certainly welfare reducing.

3.1.1 Prevention is non-contractible ($p^e = 0$)

Suppose that $p^e = 0$ and \widehat{p}^{gp} and \widehat{p}^h are set at the first best level, as described in the previous section. Then, preventive effort is too low, i.e. below the first best level \widehat{e} , if the

following holds:

$$\begin{aligned} & \left[G - \left(\int_{\underline{s}}^{\hat{z}} b^{gp}(s)f(s) + \int_{\hat{z}}^{\bar{s}} b^h(s)f(s) \right) \right] (1 - \alpha^{gp}) \\ & > \int_{\underline{s}}^{\hat{z}} c^{gp}(s)f(s)ds + (\hat{p}^{gp} + k) \int_{\hat{z}}^{\bar{s}} f(s)ds. \end{aligned} \quad (21)$$

Prevention is too low if G (utility when healthy) is sufficiently high or if altruism is sufficiently low.

If $p^e = 0$, then the optimal second-best price defined with \tilde{p}^{gp} will have to be designed in a way that it trades off optimal prevention with optimal referrals. Recall that a higher price for referrals increases preventive effort. If preventive effort is too low, i.e. $e(p^e = 0, \hat{p}^{gp}, \hat{p}^h) < e(\hat{p}^e, \hat{p}^{gp}, \hat{p}^h)$, the then optimal price under such constrained environment is distorted upwards so that $\tilde{p}^{gp}(p^e = 0) > \hat{p}^{gp}(p^e = \hat{p}^e)$.

Note that variations in the price of referrals p^{gp} have only a direct effect on preventive effort and no indirect effect through the cut-off point z^{gp} , since as already shown $\partial^2 U^{gp}(z^{gp}, e) / \partial e \partial z^{gp} = 0$. Therefore, starting at $p^{gp} = \hat{p}^{gp}$ a marginal increase in price will increase the severity cut-off point, which has no welfare effects, and increases preventive effort, which increases welfare. The opposite holds if the level of prevention is too high at $p^e = 0$, i.e. $e(p^e = 0, \hat{p}^{gp}, \hat{p}^h) > e(\hat{p}^e, \hat{p}^{gp}, \hat{p}^h)$. In this case, it is optimal to distort the referral price downwards to discourage prevention, so that $\tilde{p}^{gp}(p^e = 0) < \hat{p}^{gp}(p^e = \hat{p}^e)$.

3.1.2 Average-cost pricing with $T^h = 0$, $p^h = p^{gp}$

We now discuss the case where the hospital receives no fixed-budget component ($T^h = 0$), the price for each referral is identical to the price received by the hospital ($p^h = p^{gp}$), and the price is determined by an average-cost pricing rule so that:

$$\bar{p}^h = \frac{\int_{\underline{s}}^{\bar{s}} c^h(s)f(s)ds}{1 - F(\bar{z})} \quad (22)$$

where $\bar{z} \in [\hat{z}, z_2]$. We treat \bar{z} as exogenous. In reality this is determined by past hospital severity, as DRG prices are updated with a lag of one to three years, depending on the country.

Then, the number of referrals is too low if:

$$\bar{p}^h = \frac{\int_{\bar{z}}^{\bar{s}} c^h(s) f(s) ds}{1 - F(\bar{z})} > c^h(\hat{z}) + [b^h(\hat{z}) - b^{gp}(\hat{z})] \left(\frac{\alpha^{gp}}{\gamma^{gp}} - \frac{1}{1 + \lambda} \right)$$

Then, $\bar{p}^h > c^h(\hat{z})$ as the average cost of treatment is always higher than the cost of treatment of the patients with lowest severity. It is straightforward to show that the hospital price is too high whenever $\frac{\alpha^{gp}}{\gamma^{gp}} - \frac{1}{1 + \lambda} < 0$, ie whenever the level of altruism is sufficiently low. Even if this condition does not hold, the number of referrals is too low as long as: $\bar{p}^h - c^h(\hat{z}) > [b^h(\hat{z}) - b^{gp}(\hat{z})] \left(\frac{\alpha^{gp}}{\gamma^{gp}} - \frac{1}{1 + \lambda} \right)$.

3.1.3 Fixed budgets

Finally, we consider the case where the hospital is paid according to a fixed budget. If there is no Practice Based Commissioning, then the number of referrals and treatment is excessively low compared to the first best.

We might intuitively expect a negative referral price (ie the GP receives a price for every referral, rather than paying it) to increase the number of referrals towards the optimal level. However, this is not the case. Since the GP anticipates that the hospital will send the patients with severity below z_2 back to the GP, the number of referrals is unaffected by variations in the price of referrals. For the same reason, a positive referral price will also have no effect on GP referrals.

4 Conclusions

We have investigated the incentives which arise from introducing budgets for general practitioners (also known in England as Practice Based Commissioning) both in terms

of number of referrals and preventive effort. Our main results suggest that if hospitals are paid according to an activity-based funding mechanism (of the DRG pricing type) the introduction of PBC reduces the number of referrals. We also show, that in the absence of a price which incentivises preventive effort, the optimal price of referral is distorted upwards. In contrast, if hospitals are paid according to a fixed budget the introduction of PBC has no effect on the number of referrals.

Possible extensions of the model may include the following: i) if practice budgets for GPs are determined by past referrals and hospital utilisation then the incentive effect of PBC on preventive effort will be diluted since fewer referrals will reduce future budgets; ii) the potential surplus that GPs obtain from PBC can be spent only on non-hospital activities (eg physiotherapy services) which increase patients' benefit and which could also be funded directly by the practice thereby reducing GPs' income.

References

- [1] Barros, P.P, Martinez-Giralt, X., 2003, "Preventive care and payment systems", *Topics in Economic Analysis and Policy*, 3(1), article 10, 1-21.
- [2] Brekke, K.R., Nuscheler R., Straume O.R., 2007, Gatekeeping in health care, *Journal of Health Economics*, 26(1), 149-170.
- [3] Chalkley M., Malcomson J.M. (1998a), "Contracting for Health Services with unmonitored Quality", *The Economic Journal*, 108, 1093-1110.
- [4] Chalkley M., Malcomson J.M. (1998b), "Contracting for Health Services when patient demand does not reflect quality", *Journal of Health Economics*, 17, 1-19.
- [5] De Fraja G. (2000), "Contracts for Health Care and Asymmetric Information", *Journal of Health Economics*, 19(5), 663-677.
- [6] Dranove D. (1987), "Rate-setting by DRG and hospital specialization", *Rand Journal of Economics*, 18(3), 417-27.
- [7] Dusheiko, M., Gravelle H., Jacobs R., Smith P.C., 2006, "The effect of financial incentives on gatekeeping doctors: Evidence from a natural experiment", *Journal of Health Economics*, 25(3), 449-478.
- [8] Ellis R. P., McGuire T.G. (1986), "Provider Behaviour under prospective reimbursement: cost sharing and supply", *Journal of Health Economics*, 5, 129-51.
- [9] Ellis R. P. (1998), "Creaming, skimping and dumping: provider competition on intensive and extensive margins", *Journal of Health Economics*, 17(5), 537-55.
- [10] Hafsteinsdottir, E.J.G., Siciliani, L., 2009, DRG prospective payment systems: refine or not refine?, *Health Economics*, forthcoming.
- [11] Hammond, P., (1987). Altruism. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), *The New Palgrave: A Dictionary of Economics*. Macmillan, London, 85-87.

- [12] Jack, W., (2005), "Purchasing health care services from providers with unknown altruism", *Journal of Health Economics*, 24(1), 73-93.
- [13] Ma A.C. (1994), "Health care payment systems: cost and quality incentives", *Journal of Economics and Management Strategy*, 3(1), 93-112.
- [14] Malcomson, J.M. (2004). "Health Service Gatekeepers," *RAND Journal of Economics*, The RAND Corporation, 35(2), 401-421.
- [15] Malcomson, J.M. (2005). "Supplier Discretion Over Provision: Theory and an Application to Medical Care," *RAND Journal of Economics*, The RAND Corporation, 36(2), 412-429.
- [16] Mossialos E., (2002), "Funding health care: options for Europe", Buckingham: Open University Press, European Observatory on Health Care Systems series.
- [17] Rickman N., McGuire A., (1999), "Providers' reimbursement in the health care market", *Scottish Journal of Political Economy*, 46(1), 53-71.
- [18] Shleifer A. (1985), "A Theory of Yardstick Competition", *RAND Journal of Economics*, The RAND Corporation, 16(3), 319-327.
- [19] Siciliani, L., (2006), "Selection of treatment under prospective payment systems in the hospital sector", *Journal of Health Economics*, 25(3), 479-499.

Figure 1. Cost of hospital and GP care

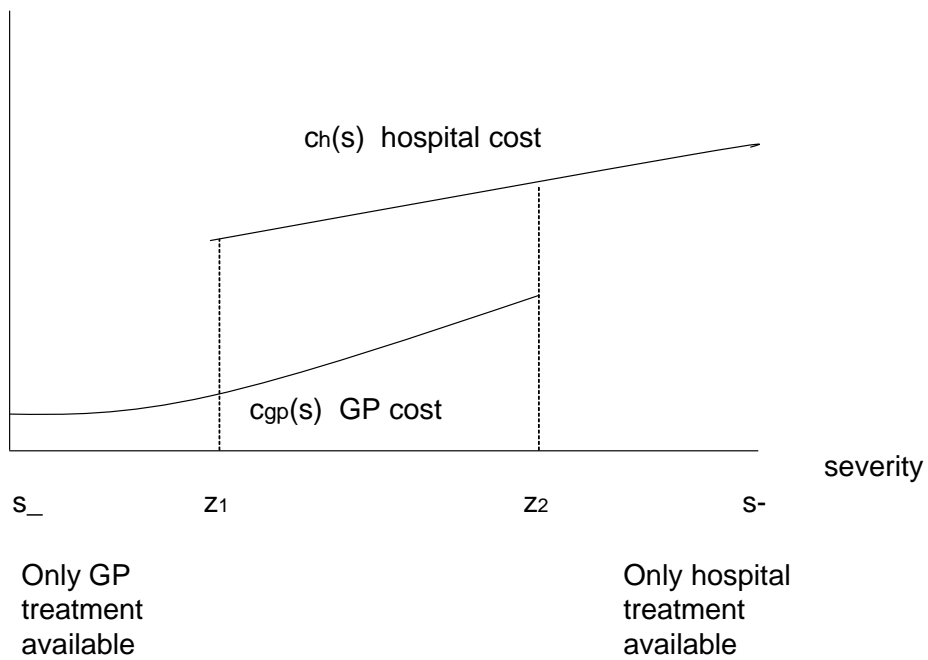


Figure 2. First best

