Felder, Stefan; Amann, Erwin

Conference Paper

No Crowding Out despite Kickbacks: Competition between Gatekeeping GPs

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2017: Alternative Geld- und Finanzarchitekturen - Session: Health Economics II, No. B11-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Felder, Stefan; Amann, Erwin (2017) : No Crowding Out despite Kickbacks: Competition between Gatekeeping GPs, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2017: Alternative Geld- und Finanzarchitekturen - Session: Health Economics II, No. B11-V2, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/168116

No Crowding Out despite Kickbacks: Competition between Gatekeeping GPs*

Erwin Amann, University of Duisburg-Essen

Stefan Felder+, University of Basel

February 23, 2017

Abstract

In health service markets, patients often rely on the advice of their general practitioner (GP) to decide which treatment best fits their needs. Specialists and hospitals, in turn, influence GPs referral decisions through kickbacks. We formulate a model with competitive heterogeneous GPs who differ in the degree to which they internalize the disutility that their patients suffer from inappropriate treatments. We prove the existence of an inefficient equilibrium that separates GPs into referrers and care providers and show that a regulator, by optimally taxing referrals and treatments, can achieve the efficient allocation if patients are rational. With naïve patients, only a second-best equilibrium is feasible.

Keywords: Health care, kickbacks, referral behavior, price competition, crowding out, second-best

JEL index: D47, D82, I11, I18, L50

---

* We thank Beat Hintermann for helpful comments.

+ Corresponding address: Faculty of Business and Economics, Department of Health Economics, Peter Merian-Weg 6, CH-4002 Basel. Email address: stefan.felder@unibas.ch.

1. Introduction

In many countries, patients have no direct access to hospital treatment or medical specialist, but need a referral from their general practitioner (GP). At the same time, price competition between health care providers and the access that they have to health care markets are often heavily regulated. As *Pauly* (1979) noticed early on, administered prices that are above marginal costs incentivize providers, who are intent on realizing potential profits, to pay kickbacks to physicians in return for patient referrals. Patients, however, might also benefit from kickbacks, since these payments foster competition between providers. In contrast, professional ethical codes and the laws of most countries prohibit referring physicians from accepting kickbacks. Kickbacks are seen as a channel exploited by profit-maximizing physicians to crowd out intrinsically motivated colleagues. However, despite prohibition and ethical reservations, kickback payments are widespread.[1]

This paper analyzes the referral behavior of GPs who are equipped with a gatekeeping role. Their patients need treatment but do not know whether they are severely ill or are only suffering from a minor illness. GPs diagnose the severity of the illness and make the referral decision. A severe illness requires a referral to inpatient care. A minor illness might also be better treated in a hospital if the cost-benefit ratio of inpatient care is sufficiently low. If not, GPs might still refer patients without treating them because of a kickback offered by the hospital. We consider heterogeneous GPs who differ in the degree they internalize the disutility their patients experience as a result of unnecessary hospital treatment. Finally, we extend the model to include heterogeneous patients who differ with respect to the disutility from inpatient care.

Kickbacks have hardly been addressed in the literature so far. *Owen* (1977) discusses the role of kickbacks that providers of title-insurance services pay to real estate brokers to steer homebuyers. *Pauly* (1979) shows that kickbacks might induce GPs to refer their patients rather than perform lower quality procedures themselves, and thus benefit their patients. *Felder* (2016) analyzes kickbacks when a medical service is a credence good and shows that an equilibrium with overcharging and overtreatment can exist, depending on the patient's capacity to verify the treatment. The present paper breaks new ground, as it considers both heterogeneous GPs and heterogeneous patients. *Inderst* and *Ottaviani* (2012) analyze Hotelling competition between two sellers through kickbacks to intermediaries who advise consumers. Our framework differs, as we consider a monopolistic hospital which provides both treatments, and which competes with GPs who can provide only the minor treatment. In fact, hospitals often have regional market power, due to high setup costs and patients' preferences for nearby inpatient care. Without market power, competition would drive inpatient prices down to opportunity costs (*Kerschbamer* and *Dulleck*, 2006), and kickbacks would not persist. Furthermore, we include price competition in the outpatient sector.

[1] In a 2012 survey of 1,141 medical providers in Germany, 49 percent of physicians agreed or partly agreed that, despite the prohibition, kickbacks are quite common, and 20 percent indicated that they are very common. Among non-medical professionals, the affirmation rate was even higher (*Bussmann*, 2012).

There is a broad literature on gatekeeping in health care. *Brekke* et al. (2007) investigate the informational role of GP gatekeepers in secondary health care markets where two hospitals compete in quality and specialization. They complement the multi-task agency literature on the economics of general practice, e.g., by *Garcia Mariñoso* and *Jelovac* (2003) and *Malcomson* (2004). Optimal payment systems are derived that induce GPs to conduct diagnoses and incentivize efficient referral and treatment decisions. The role of price competition is not captured in this literature. Instead, price regulation is present, as in *Brekke* et al. (2007), or the principal agent structure between a public insurer and a GP under asymmetric information is highlighted.

The following Section 2 presents the basic model with ex ante homogeneous patients and heterogeneous GPs who differ in their degree of altruistic preferences. We consider a monopolistic hospital that chooses the profit-maximizing kickback, given administered treatment and referral prices in the outpatient sector. While inpatient care involves a specific disutility to patients, we assume that the hospital has a comparative cost advantage when treating patients suffering from a minor problem. The cost-benefit ratio of inpatient care, then, determines in which of these sectors the patients with minor illnesses should be treated. By setting the appropriate referral and treatment prices in the outpatient sector, the regulator can achieve the efficient allocation. If he forbids kickbacks and is able to enforce this ruling, the efficient allocation can again be achieved by regulating outpatient prices.

Section 3 investigates the effect of price competition between GPs. We differentiate equilibria with naïve and rational expectations. While naïve patients expect that their GP will act in their best interest, rational patients foresee that their GP might unnecessarily refer them to the hospital. Depending on the cost-benefit ratio of treating patients with a minor illness in hospital, an efficient crowding-out equilibrium with naïve patients or a rational pooling equilibrium arises where also intrinsically motivated GPs remain in the outpatient sector. This pooling equilibrium is, however, inefficient. We then show that the regulator can produce the efficient allocation by subsidizing referrals or taxing treatments in the outpatient sector, whether kickbacks are forbidden or not.

Section 4 introduces heterogeneous patients who differ in the degree they suffer from the treatment administered in the hospital. In this environment, either a crowding-out equilibrium or an equilibrium exists that separates GPs into referrers and care providers. Optimally set taxes can produce a second-best equilibrium, irrespective of whether kickbacks are forbidden and enforceable or not. Section 5 discusses the results and concludes.

2. Administered outpatient prices for treatment and referrals

We assume a primary care market with administered prices, and a monopolized secondary care market, and consider the health service to be a credence good. The monopolistic hospital sets the profit-maximizing price and chooses the optimal kickback it pays to general practitioners (GP) for

their referral of patients.[2] Patients seek treatment for their illnesses. They do not know whether they are severely ill or only suffering from a minor medical problem. When their health is restored through treatment, they uniformly achieve utility *r*, which represents their reservation price for medical treatment. When comparing inpatient with outpatient care, patients who are treated in hospital experience a disutility equal to *s*, which reduces their utility from inpatient treatment to $r-s$. Patients have no direct access to inpatient care. They visit their GP who acts as a gatekeeper. He diagnoses the patient and decides whether to treat him or refer him to inpatient care. Severely ill patients will always be referred to the hospital and receive inpatient treatment. GPs receive a kickback payment $\kappa \geq 0$ from the hospital for referring patients with a minor illness. We assume that GPs are heterogeneous with respect to their minor ill patients' disutility from inpatient care. A GP of type α internalizes a share α of monetized patient disutility *s*. But GPs are also imperfect agents in the sense that they misreport the diagnosis to patients who have minor illnesses in order to refer them to inpatient care.[3] GP types are drawn independently from the cumulative distribution function $F(\alpha)$ with a differentiable strictly positive density $f(\alpha)$ on the interval $[0,1]$. $F(\alpha)$ is common knowledge, but a GP's type α is his private information. The diagnosis cost is *d* and μ denotes the population share of patients with a minor medical problem. Referral price p_r and treatment price p_t are set by the regulator, while the hospital chooses its treatment price p_H and the kickback payment κ.

Faced with p_r, p_t, κ and the cost of outpatient treatment $\bar{c}_G$, GPs determine the share *x* of patients with a minor problem whom they will refer to the hospital, by maximizing their profit:

$$\max_x \pi_{GP} = (1-\mu)(p_r - d) + \mu\left[(1-x)(p_t - \bar{c}_G - d) + x(\kappa + p_r - d - \alpha s)\right]. \tag{1}$$

A GP of type α will refer patients with minor problems to inpatient care, provided that

$$\kappa + p_r - d - \alpha s \geq p_t - \bar{c}_G - d\,. \tag{2}$$

With heterogeneous GPs, we obtain the pivotal GP $\tilde{\alpha}$ who is indifferent between referral and providing care:

$$\tilde{\alpha} = \frac{\kappa - (p_t - p_r - \bar{c}_G)}{s}, \tag{3}$$

where $p_t - p_r - \bar{c}_G$ is the GPs' net profit from treating the patient, rather than referring him. GPs with $\alpha \leq \tilde{\alpha}$ will refer patients suffering from a minor medical problem (i.e., $x = 1$), while GPs

[2] Note that the hospital price might also be an administered price. This would not change our analysis as the main variable of interest is the kickback payment.

[3] The physician agency literature analyzes in detail strategic reasons for GPs to make false reports (for an overview, see *McGuire*, 2000).

with $\alpha > \tilde{\alpha}$ will treat them (i.e., $x = 0$). An equilibrium with $\tilde{\alpha} > 0$ implies that a referral is basically profitable: $\kappa + p_r > p_t - \bar{c}_G$. Thus, GPs who are mostly extrinsically motivated will refer patients with minor illnesses to inpatient care, even though the latter do not need this treatment. For GP types $\alpha > \tilde{\alpha}$, the profit from a referral, taking into account the patient's disutility, is smaller than the profit from treatment in the GP's practice. Hence, GPs with $\alpha > \tilde{\alpha}$ will treat patients with minor illnesses.

If the regulator sets outpatient prices at marginal costs, $p_t = c_t + d$ and $p_r = d$, (3) becomes $\tilde{\alpha} = \kappa / s$. In this case, the kickback payment equals the disutility from hospital treatment for patients with a minor illness as perceived by the pivotal GPs $\tilde{\alpha}$.

Consider, then, the monopolistic hospital H. Assume that c_H and $\underline{c}_G$ are the respective costs for the inpatient treatment of major and minor cases, with $c_H > \underline{c}_G$. The hospital does not incur a diagnostic cost, as it is assumed that it knows the patients' diagnoses from the referring GPs. It sets the uniform profit maximizing price $p_H = r - s$. Since only GPs with $\alpha \leq \tilde{\alpha}$ refer patients with a minor illness to outpatient care, the monopolistic hospital's profit maximization problem can be written as follows:

$$\max_{\kappa} \pi_H = (1-\mu)(p_H - c_H) + \mu F(\tilde{\alpha})(p_H - \kappa - \underline{c}_G). \tag{4}$$

The FOC for the profit-maximizing kickback reads

$$\frac{d\pi_H}{d\kappa} = \mu \frac{dF(\tilde{\alpha})}{d\kappa}\left(p_H - \kappa^* - \underline{c}_G\right) - \mu F(\tilde{\alpha}) = 0. \tag{5}$$

With $\frac{dF(\tilde{\alpha})}{d\kappa} = \frac{\partial F(\tilde{\alpha})}{\partial \tilde{\alpha}} \frac{d\tilde{\alpha}}{d\kappa} = f(\tilde{\alpha}) \cdot \frac{1}{s}$ (see (3)), the condition for maximum profit satisfies:

$$\frac{p_H - \kappa^* + \underline{c}_G}{s} f(\tilde{\alpha}) = F(\tilde{\alpha}), \tag{6}$$

where the LHS indicates the profit per patient at the intensive margin, while the RHS equals the additional kickback payments for infra-marginal referrals from a marginal increase in the kickback.

Solving for optimal kickbacks gives

$$\kappa^* = p_H - \underline{c}_G - s\frac{F(\tilde{\alpha})}{f(\tilde{\alpha})}. \tag{7}$$

For α uniformly distributed in the interval $[0,1]$, $F(\alpha)=\alpha$ and $f(\alpha)=1$ hold. Using (3), we obtain

$$\kappa^{*}=\frac{(p_H-\underline{c}_G)+(p_t-p_r-\overline{c}_G)}{2}. \tag{8}$$

In this case, the profit-maximizing kickback equals the average net profits that the hospital and the GP realize from treating a patient who has a minor medical problem. The hospital's profit is $\tilde{\alpha}s$. Thus, when it sets the kickback, the hospital exploits the disutility from unnecessary inpatient treatment of patients as perceived by the pivotal GPs. It lowers the kickback and attains a positive profit for the treatment of minor cases. With outpatient prices at marginal costs, we obtain $\kappa^{*}=(p_H-\underline{c}_G)/2$ and for the pivotal GPs $\tilde{\alpha}=(p_H-\underline{c}_G)/2s$, provided that $(p_H-\underline{c}_G)/2s\le 1$.

Assume that, while inpatient care involves the disutility *s*, the cost of inpatient care for patients suffering from a minor illness is lower than in the outpatient setting: $\Delta c_G=\overline{c}_G-\underline{c}_G>0$. This implies that, depending on the disutility of inpatient care *s*, the society may benefit if the hospital also treats minor cases.

Social welfare[4] is defined as follows:

$$W=V+\pi+\mu s\int_0^{\tilde{\alpha}}\alpha f(\alpha)\,d\alpha. \tag{9}$$

The consumer surplus *V* amounts to

$$V=(1-\mu)(r-s-p_H-p_r)+\mu\big(F(\tilde{\alpha})(r-s-p_H-p_r)+(1-F(\tilde{\alpha}))(r-p_t)\big) \tag{10}$$

and for the profit π, we obtain

$$\begin{aligned}\pi=\pi_{GP}+\pi_H&=(1-\mu)(p_r-d+p_H-c_H)\\&+\mu\left((1-F(\tilde{\alpha}))(p_t-\overline{c}_G-d)+F(\tilde{\alpha})(\kappa+p_r-d+p_H-\kappa-\underline{c}_G)-s\int_0^{\tilde{\alpha}}\alpha f(\alpha)\,d\alpha\right).\end{aligned} \tag{11}$$

Social welfare, then, becomes

$$\begin{aligned}W&=r-d-(1-\mu)c_H-\mu\big(F(\tilde{\alpha})\underline{c}_G+(1-F(\tilde{\alpha}))\overline{c}_G\big)-\big(1-\mu(1-F(\tilde{\alpha}))\big)s\\&=r-d-(1-\mu)c_H-\mu\overline{c}_G-(1-\mu)s+\mu F(\tilde{\alpha})(\Delta c_G-s).\end{aligned} \tag{12}$$

[4] This definition avoids double counting of the GPs' evaluation of the patient's disutility from unnecessary treatment in the hospital. On double counting, see *Chalkley* and *Malcomson* (1998) in a physician reimbursement context, and, more generally, *Ng* (1983).

Welfare per patient equals the utility from restoring health *r* minus the cost of diagnosis *d* minus average treatment cost, $(1-\mu)c_H + \mu\left(F(\tilde{\alpha})\underline{c}_G + (1-F(\tilde{\alpha}))\bar{c}_G\right)$, minus the disutility of inpatient care averaged over all patients, $\left(1-\mu(1-F(\tilde{\alpha}))\right)s$.

An efficient allocation requires for $s \le \Delta c_G$, that all patients should be treated by the hospital and, by comparison, for $s > \Delta c_G$, that it is optimal to treat all patients who suffer from a minor illness in the outpatient setting (see the second line of (12)).

From a societal perspective, the relative size of disutility and cost advantage of treating a patient who has a minor illness is decisive. To induce the efficient allocation, the regulator can set the appropriate treatment and referral prices. He might also consider forbidding kickback payments if such a ban can be enforced. Assume that the regulator knows the disutility of inpatient treatment *s*, the diagnostic and the treatment costs in the out- and inpatient sectors, including the hospital's cost advantage when treating a patient who has a minor illness Δc_G.

Proposition 1: By setting the optimal referral and treatment prices, p_r and p_t, the regulator can induce an efficient allocation.

Proof: If $s < \Delta c_G$, the efficient allocation requires $\tilde{\alpha} = 1$. The regulator sets $p_t = 0$ and $p_r = d + s$. Thus, treatment is not profitable and even GPs with $\alpha = 1$ are willing to refer their patients to the hospital. The optimal kickback is zero, as the hospital receives all patients anyway. If $s > \Delta c_G$, three conditions are required to obtain $\tilde{\alpha} = 0$: (i) GPs have no incentive to refer: $p_t - p_r - \bar{c}_G > \kappa$; (ii) the kickback is at its upper bound: $\kappa^{max} = p_H - \underline{c}_G$; and (iii) outpatient treatment renders a non-negative profit: $\pi_{GP} = -d + (1-\mu)p_r + \mu(p_t - \bar{c}_G) \ge 0$. Combining these three conditions, we obtain $p_t \ge \bar{c}_G + d + (1-\mu)(p_H - \underline{c}_G)$ and $p_r < d - \mu(p_H - \underline{c}_G)$. ■

Optimally set referral and treatment prices can also produce the efficient allocation if a ban on kickbacks can be enforced. For $s > \Delta c_G$, the regulator sets $p_r < d$ and $p_t \ge \bar{c}_G + d$, while for $s < \Delta c_G$, in order to avoid treatment in the outpatient sector, he sets $p_r \ge d + s$ and $p_t < d + \bar{c}_G - (1-\mu)s$.

Note that in order to achieve the efficient allocation, the regulator generally deviates from marginal cost prices.

3. Competition between GPs

Consider, first, the case with naïve patients in the sense that they do not expect to be referred to the hospital if they are only suffering from a minor illness. They will minimize their expected expenditure $\mu p_t + (1-\mu)(p_r + p_H)$ and, thus, choose the GP with the lowest treatment price p_t.

Proposition 2 (naïve patients): In the equilibrium with naïve patients, only GPs with $\alpha = 0$ survive in the market, where their profit is zero. The equilibrium kickback is $\kappa^* = 0$.

Proof: With given kickback κ, the pivotal GP $\tilde{\alpha}$ separates the GPs who refer from those who do not (see (3)). For those GPs who only refer their patients ($\alpha \le \tilde{\alpha}$), the profit $\pi_{GP} = p_r - d + \mu(\kappa - \alpha s)$ is independent of the treatment price. For those GPs who refer severely ill patients to the hospital, but who treat patients with minor medical problems in their practice, we obtain $\pi_{GP} = (1-\mu)(p_r - d) + \mu(p_t - d - \bar{c}_G)$. Now, for all $p_t < d + \bar{c}_G$, each GP with $\alpha \le \tilde{\alpha}$ can crowd out GPs with $\alpha > \tilde{\alpha}$. Hence, only GPs with $\alpha \le \tilde{\alpha}$ remain in the market. The minimum price $p_r = d - \mu(\kappa - \alpha l)$ that induces referrals is increasing in α. Thus, a negative GP selection with regard to type α takes place, and only the extrinsically motivated GPs $\alpha = 0$ survive. Their profit is zero, irrespective of the size of κ. The monopolistic hospital will choose the minimum kickback, since with $\kappa^* = 0$, its profits are maximized: $\pi_H^* = r - p_H - (1-\mu)c_H - \mu \underline{c}_G$. ■

As an alternative scenario, assume that patients expect a referral even if they only have a minor medical problem. An equilibrium might then exist that separates GPs into "referrers" and "care providers". Referrers *R* do not treat patients: ($p_r^R > 0$, no treatment). Care providers *C's* prices are split furthest apart; i.e., they post the price pair $(p_r^C = 0, p_t^C > 0)$. The preference parameter α is relevant to the extent that the pivotal GPs separate the two groups, while α is irrelevant for the competitive treatment price. We assume that for GPs in group *R* $\alpha = 0$ holds, as these GPs' prices signal to patients that they will always refer.[5]

Proposition 3 (rational patients):

i) For $s < \Delta c_G$, a crowding-out equilibrium exists where all patients are referred and the referral price is non-positive. This equilibrium is efficient.

ii) For $s > \Delta c_G$, no equilibrium exists that separates "referrers" from "care providers".

[5] This assumption simplifies the analysis, as it implies that GPs' profits in group *R* do not vary with α. If they did, it would lead to crowding out among the members of group *R*.

iii) For $s > \Delta c_G$, a pooling equilibrium exists where both referrers and care providers are in the outpatient sector, with $\tilde{\alpha} = \left(r - s - \underline{c}_G - d\left(1-\mu\right)/\mu\right)/2s$ as the pivotal GPs. This allocation is not efficient.

Proof: see Appendix.

The separating equilibrium for $s > \Delta c_G$ does not exist, since the kickback is too low for the referrers to offer a competitive price. In the pooling equilibrium, GPs of type $\alpha < \tilde{\alpha}$ will falsely report to their patients who have minor illnesses as being severely ill and need a referral to receive inpatient care. The inefficient pooling equilibrium will only emerge if $s > \Delta c_G$. For $s < \Delta c_G$, the crowding-out equilibrium exists, where the hospital pays no kickbacks ($\kappa^* = 0$) and only the extrinsic GPs remain in the market.

With competition, the regulator can no longer control prices. However, he might tax or subsidize referrals and treatments in the outpatient sector.[6] Denote τ_r and τ_t as the corresponding tax rates.

Proposition 4 (rational patients): A tax on treatments in the outpatient sector exists that leads to the efficient allocation.

Proof: If $s < \Delta c_G$, the regulator will be inactive. If $s > \Delta c_G$, he imposes a sufficiently high tax on referrals such that GPs abstain from making referrals. For referrers, the profit is $\pi_{GP} = p_r - \tau_r - d + \mu\left(\kappa - \alpha s\right)$. Altruistically motivated GPs have a lower incentive to refer minor cases. So the regulator has to remove the referral incentive for the extrinsically motivated GPs. This implies that $\tau_r = p_r + \mu\kappa - d$. Competition between referrers lowers their price down to marginal costs, so that $\tau_r = \mu\kappa$. A referral tax $\tau_r > \mu\kappa$ will, then, force referrers out of the market. ■

If it is possible to enforce the prohibition of kickbacks, such a regulation would produce the efficient allocation for $s > \Delta c_G$. For $s < \Delta c_G$, provided that $p_t - \bar{c}_G - d < p_r - \tau_r - d - s$ holds, GPs with $\alpha = 1$ will also refer their patients. With the marginal cost price for treatment $p_t = \bar{c}_G + d$, a referral subsidy of $\tau_r < -s$ will lead to the efficient allocation.

[6] The regulator can either tax treatments or subsidize referrals. However, it suffices to employ only one tax instrument.

4. Heterogeneous patients

Assume that patients are heterogeneous with regard to the disutility of inpatient care: $s \in [0, \bar{s}]$, where the cumulative distribution function $G(s)$ and a strictly positive density $g(s)$ are known to both the hospital and the regulator. We assume that $\bar{s} > s^e = \Delta c_G$; i.e., from a societal perspective, patients with minor illnesses of type $s \le s^e$ should receive inpatient care, while type $s > s^e$ patients should not.

Proposition 5 (rational patients): An equilibrium exists that separates "referrers" from "care providers" if patients are heterogeneous with regard to *s* and have rational expectations. This equilibrium is not efficient.

Proof: A patient's utility *s* from visiting a GP belonging to group *R* is $u^R(s) = r - s - p_H - p_r^R$, while his utility from visiting a GP belonging to group *C* is $u^C(s) = (1-\mu)(r - s - p_H) + \mu(r - p_t^C)$. For the pivotal patients who are indifferent with regard to the type of physician they decide to consult, we obtain $\tilde{s} = p_t^C - p_H - p_r^R / \mu$. Competitive prices in the outpatient sector are $p_r^R = d - \mu(\kappa - \alpha s)$ and $p_t^C = d/\mu + \bar{c}_G$. Let us again assume that GPs of group *R* will not internalize their patients' disutility from inpatient treatment. This implies that $p_r^R = d - \mu\kappa$ and that for the pivotal patients $\tilde{s} = \bar{c}_G - p_H + \kappa$. The hospital maximizes the profit from treating minor cases: $\mu G(\tilde{s})(p_H - \kappa - \underline{c}_G)$, with $\tilde{s} = \bar{c}_G - p_H + \kappa$. This implies that the optimal kickback is $\kappa^* = p_H - \underline{c}_G - G(\tilde{s})/g(\tilde{s})$ (note the analogy to (7)) and that for the marginal patients $\tilde{s}^* = \Delta c_G - G(\tilde{s}^*)/g(\tilde{s}^*)$. As $\tilde{s}^* < s^e = \Delta c_G$, too few patients are referred. ■

The two groups are separated by the marginal GPs within group *C* for whom $\tilde{\alpha}^* \tilde{s}^* = \kappa^* - (p_t^C - p_t^R - \bar{c}_G)$ holds. With competitive prices, we obtain $\tilde{\alpha}^* = (1-\mu)(\kappa^* - d/\mu)/\tilde{s}^*$. The share of GPs who refer becomes $F(\tilde{\alpha}^*) = (1-\mu)(\mu\kappa^* - d)/\tilde{s}^*$. For a uniformly distributed *s*, one finds $G(\tilde{s})/g(\tilde{s}) = \tilde{s}$, $\kappa^* = p_H - \Delta c_G/2$ and $\tilde{s}^* = \Delta c_G/2$.

Referrals are, thus, too low. With a uniformly distributed disutility of hospital care, only half of the patients requiring a referral are actually referred. This allocation is, nevertheless, better than that under a kickback ban where no referrals would take place.

Proposition 6 (rational patients): The efficient allocation emerges if referrals are subsidized with the rate $\tau_r^* = -\mu G(\tilde{s})/g(\tilde{s})$.

Proof: For the pivotal patient, it holds that $\tilde{s} = p_t^C - p_H - \left(p_r^R + \tau_r\right)/\mu$. Taking into account the zero-profit condition, we obtain $\tilde{s} = \overline{c}_G - p_H - \tau_r/\mu + \kappa$. The profit-maximizing hospital chooses $\kappa^* = p_H - \underline{c}_G - \tau_r/\mu - G\left(\tilde{s}\right)/g\left(\tilde{s}\right)$, so that $\tilde{s}^* = \Delta c_G - \tau_r/\mu - G\left(\tilde{s}\right)/g\left(\tilde{s}\right)$ arises. The regulator sets $\tau_r^* = -\mu G\left(\tilde{s}\right)/g\left(\tilde{s}\right)$ and obtains $\tilde{s}^* = s^e$. ■

Fig. 1 illustrates the inefficient and the efficient allocations. The utility difference for the patients visiting the referrers amounts to $\Delta u^R = \mu\left(\kappa^*\left(\tau_r^*\right) - \kappa^*\right)$. The difference in the kickback payment equals the change in the referral price, as the latter falls by the full subsidy rate because of constant marginal costs and pure competition. This change must be positive, as the kickback is at its upper bound if the optimal subsidy applies. Thus, patients visiting the referrers benefit from the subsidy. For the patients consulting the care providers, the utility increases by $\Delta u^C = (1-\mu)\left(\kappa^*\left(\tau_r^*\right) - \kappa^*\right)$, since they too benefit from the decrease in the referral price.

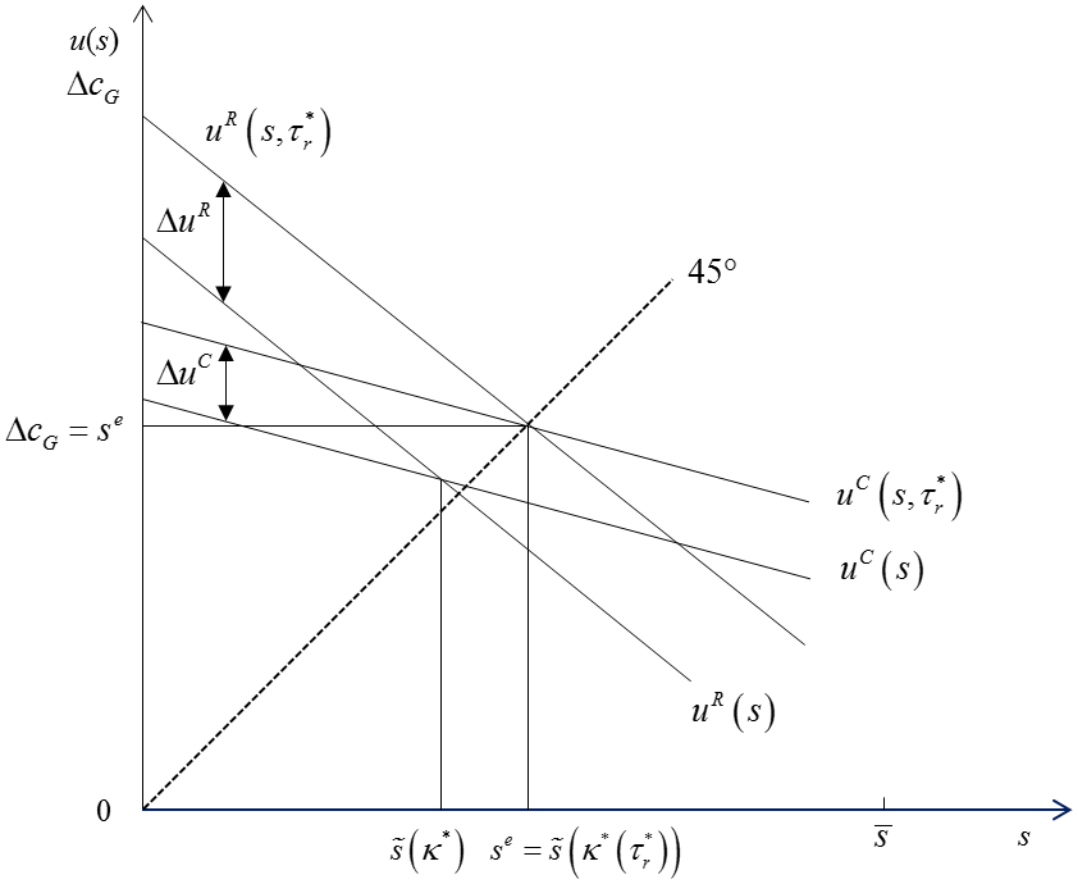

Figure 1: The efficient allocation with subsidy rate: $\tau_r^* = -\mu G\left(s^e\right)/g\left(s^e\right)$

If no subsidy applies, $\tilde{s}\left(\kappa^*\right)$ marks the marginal patient who chooses the outpatient care provider. The optimal subsidy $\tau_r^* = -\mu G\left(\tilde{s}\right)/g\left(\tilde{s}\right)$ lowers the referral price such that $\tilde{s}\left(\kappa^*\left(\tau_R^*\right)\right) = s^e$ holds. This allocation is efficient, since $s^e = \Delta c_G$ and all patients with $s < s^e$ visit a referrer and receive inpatient care. For *s* uniformly distributed, the optimal subsidy rate amounts to $\tau_r^* = -\mu\Delta c_G$.

Note that with an optimal subsidy, the kickback reaches its upper bound: $\kappa^* = p_H - \underline{c}_G$. The regulator lowers the referring price such that all patients of type $s \le \Delta c_G$ visit referring GPs. The hospital receives zero-profit for treating patients with minor illnesses, while the patients benefit from the lower price they pay to the referrer.

If patients are naïve or do not know their *s*, an efficient allocation is no longer feasible.

Proposition 7 (naïve patients): Price competition between heterogeneous GPs and heterogeneous patients seeking medical care leads to heterogeneous treatment. Patients with small *s* are more likely to be referred, while patients with high *s* are treated by a GP. This equilibrium is not efficient.

Proof: The marginal patient of type *s* who receives treatment by a GP of type α is determined by $p_t - \bar{c}_G - d = \kappa + p_r - d - \alpha s$. The marginal patient is thus $s(\alpha) = \left(\kappa + p_r - p_t + \bar{c}_G\right)/\alpha$. ■

In order to analyze the potential for a second-best equilibrium, we assume a uniform distribution with respect to both *s* and α. For technical reasons, we further assume that the GPs' preference parameter α has a strictly positive lower bound: $\alpha \in \left[\underline{\alpha},1\right]$ with $\underline{\alpha} \gg 0$. Finally, we assume that $\bar{s} > \left(\kappa - \left(p_t - p_r - \bar{c}_G\right)\right)/\underline{\alpha}$, which ensures that all GPs, including the least altruistically motivated GPs, will not refer all of their patients to inpatient care. A GP of type α refers patients of type *s*, provided that $\alpha s \le \kappa - \left(p_t - p_r - \bar{c}_G\right)$. The share of patients with a minor health problem who are referred to the hospital, then, amounts to

$$x = \int_{\underline{\alpha}}^{1} \frac{\kappa - \left(p_t - p_r - \bar{c}_G\right)}{\alpha} f(\alpha)\, d\alpha = \frac{p_t - p_r - \bar{c}_G - \kappa}{1 - \underline{\alpha}} \ln \underline{\alpha}\,. \tag{13}$$

With regard to the profit-maximizing kickback, the monopolistic hospital will solve the maximum problem

$$\max_{\kappa} \pi_H = (1-\mu)\left(p_H - c_H\right) + \mu x \left(p_H - \kappa - \underline{c}_G\right). \tag{14}$$

From the FOC and (13), given that $dx/d\kappa = -\ln\underline{\alpha}/(1-\underline{\alpha})$, we obtain for the optimal kickback $\kappa^* = (p_H - \underline{c}_G + p_t - p_r - \bar{c}_G)/2$, which is equivalent to (8). With marginal cost prices in outpatient care, the optimal kickback becomes $\kappa^* = (p_H - \underline{c}_G)/2$.

All GPs refer those patients whose disutility from inpatient care is low, but depending on the altruistic preference parameter α, to a greater or lesser extent. Efficiency requires that only those patients are referred for whom $s \le \Delta c_G$ holds. This will be the case for marginal cost prices, provided that

$$\alpha = \frac{\kappa^*}{\bar{c}_G - \underline{c}_G} = \frac{p_H - \underline{c}_G}{2\Delta c_G}. \tag{15}$$

It follows that GPs with $\alpha^e = (p_H - \underline{c}_G)/2\Delta c_G$ refer an optimal number of patients, while GPs of type $\alpha < \alpha^e$ refer too many patients, and GPs with $\alpha > \hat{\alpha}$ refer too few patients. Fig. 2 illustrates. Along the inverse $s(\alpha) = (p_H - \underline{c}_G)/2\alpha$, a GP of type α is indifferent between referring and not referring a patient of type s. The area under the inverse curve indicates the share of patients x who are referred to the hospital. With marginal cost prices, we obtain $x = \ln\underline{\alpha} \cdot (p_H - \underline{c}_G)/2(1-\underline{\alpha})$. The area A reflects those patients who should not have been referred to inpatient care, while the area B represents those patients who should have been referred, but are treated in the outpatient setting instead.

Proposition 8: Competitive prices result in an inefficient allocation. Marginal cost prices are second-best if $\alpha^* = \sqrt{\underline{\alpha}}$.

Proof: An efficient allocation requires that all patients of type $s \le \Delta c_G$ are referred. According to Fig. 2, heterogeneous GPs will lead to either too many referrals (A), too few (B) or both. The efficient number of referrals would only be realized if all GPs were identical of type $\alpha = \alpha^e$.

A change in prices shifts the function $s(\alpha)$ by $\Delta p/2\alpha$. Ignoring the shift of α^e, which is of second order for the welfare effect, welfare changes according to the referral shift of infra-marginal patients: $\int_{\underline{\alpha}}^{\alpha^*} \frac{\Delta p}{2(1-\underline{\alpha})\alpha}\, d\alpha - \int_{\alpha^*}^{1} \frac{\Delta p}{2(1-\underline{\alpha})\alpha}\, d\alpha$. Positive Δp increases A and decreases B, while negative Δp go in the inverse directions. The second-best allocation, then, gives $(\ln\alpha^* - \ln\underline{\alpha})\frac{\Delta p}{2(1-\underline{\alpha})} + (\ln 1 - \ln\alpha^*)\frac{\Delta p}{2(1-\underline{\alpha})} = 0 \Leftrightarrow \ln\alpha^* = \frac{\ln\underline{\alpha}}{2} \Leftrightarrow \alpha^* = \sqrt{\underline{\alpha}}$. ■

As Proposition 8 revealed, price competition does not lead to the second-best allocation, partially owing to the hospital's behavior. The welfare effect will be unclear, once kickbacks rise to the equilibrium amount: For profit-maximizing κ, we obtain $s\left(\sqrt{\underline{\alpha}}\right)=\left(p_H-\underline{c}_G\right)/2\sqrt{\underline{\alpha}}$, which can be significantly above s^e. This offers options for the regulator to levy taxes on referrals and treatment and to consider issuing a prohibition against kickbacks in order to improve welfare.

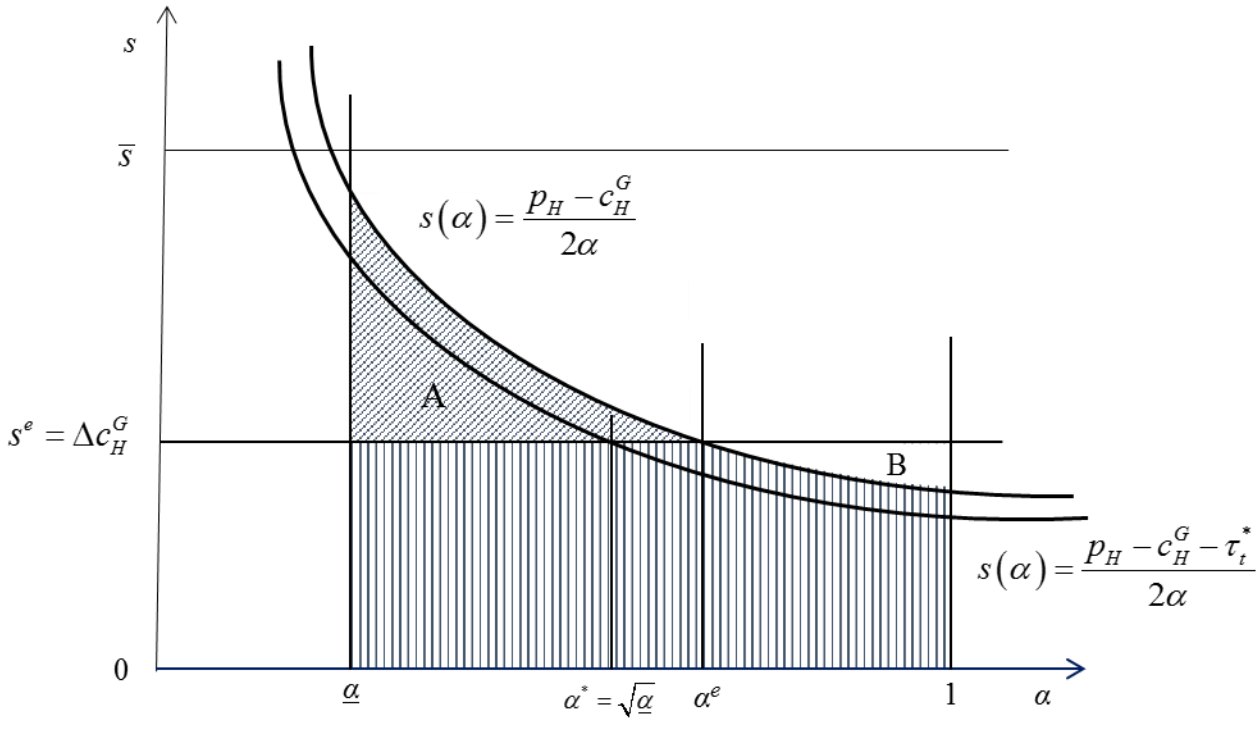

Figure 2: Misallocation of patients and the second-best allocation

Proposition 9: With uniformly distributed α and s, a treatment tax equal to $\tau_t^*=p_H-\underline{c}_G-2\sqrt{\underline{\alpha}}\Delta c_G$ leads to the second-best. If a prohibition against kickbacks can be enforced $(\kappa=0)$ and uniformly distributed α and s apply, a referral tax equal to $\tau_r^*=\mu\sqrt{\underline{\alpha}}\Delta c_G$ combined with a treatment subsidy equal to $\tau_t^*=-(1-\mu)\sqrt{\underline{\alpha}}\Delta c_G$ lead to the second-best.

Proof: With a treatment tax τ_t, the pivotal patient for a GP of type α becomes $s(\alpha)=\left(p_H-\underline{c}_G-\tau_t\right)/2\alpha$. In order to induce the critical GP of type $\alpha=\sqrt{\underline{\alpha}}$ to refer adequately, we need $s\left(\sqrt{\underline{\alpha}}\right)=\left(p_H-\underline{c}_G-\tau_t\right)/2\sqrt{\underline{\alpha}}=s^e=\Delta c_G$. The optimal tax, then, is $\tau_t^*=p_H-\underline{c}_G-2\sqrt{\underline{\alpha}}\Delta c_G$, which can be positive or negative.

If a prohibition against kickbacks can be enforced: For the pivotal GPs who treat type-*s* patients, we have $\tilde{\alpha}s = p_r + c_G - p_t$. In the second-best, $s\left(\sqrt{\underline{\alpha}}\right) = s^e = \Delta c_G$ holds. This leads to $p_t - p_r = \bar{c}_G - \sqrt{\underline{\alpha}}\Delta c_G$. From the zero profit condition, we obtain $p_r^* = d + \mu\sqrt{\underline{\alpha}}\Delta c_G$ and $p_t^* = d + \bar{c}_G - (1-\mu)\sqrt{\underline{\alpha}}\Delta c_G$. If the tax rates are set such that net prices equal marginal costs, the second-best tax rates result. ■

Fig. 2 illustrates the second-best allocation. A treatment tax for GPs decreases the hospital's kickback payment, which will shift the $s(\alpha)$ curve downwards. This, in turn, will result in fewer GP referrals. If he optimally sets the tax, the regulator can achieve the second-best.

If the hospital is free to set the kickback, the regulator can tax or subsidize treatment in the outpatient sector to achieve the second-best. If, on the other hand, kickbacks are forbidden, the regulator needs both taxes to obtain the second-best. As patients are naïve or do not know their specific disutility from inpatient care, their behavior is entirely determined by the prices. Since all GPs have the same price structure, patients are indifferent regarding their choice of GP. By taxing referrals and treatment adequately, the regulator can influence GPs' referral behavior and achieve the second-best.

5. Discussion and conclusion

This paper investigates the market outcome of kickbacks paid by a monopolistic hospital to competitive GPs in return for patient referrals. Kickbacks can incentivize defrauding behavior on the part of physicians who refer their patients to the hospital. This is to the disadvantage of patients suffering from minor illnesses, since inpatient treatment involves a specific disutility. On the other hand, kickbacks can enhance welfare, because they incentivize GPs to refer patients instead of performing a higher-cost procedure themselves. Patients need one of two possible treatments (minor or major) and differ in the disutility they experience if unnecessarily treated in the hospital.

We have assumed that patients can verify neither the diagnosis nor the treatment. Instead, one might assume that patients are able to verify the kind of treatment they receive, but not their diagnosed health status before treatment. This excludes the possibility of overcharging; i.e., where a patient who has a minor illness receives the appropriate inexpensive treatment, but pays an excessive price. The hospital's price for a minor treatment will then be bounded from above by the outpatient price. Partial verifiability, however, gives rise to overtreatment, since the hospital might have an incentive to employ the expensive treatment, which still allows it to charge the monopoly price (see *Kerschbamer* and *Dulleck*, 2006). From a societal perspective, overtreatment should be prevented, as it leads to an overuse of resources. Prohibiting kickbacks can be beneficial in this environment. However, introducing competition in the inpatient market will lead to marginal cost prices and to the disappearance of kickbacks (for details in a related model, see *Felder*, 2016).

This paper has shown the existence of an equilibrium that separates GPs into two groups: referrers, who refer all patients to the hospital irrespective of whether they are severely ill or only suffering from a minor illness, and care providers, who only refer severely ill patients. Patients select which group of GPs (referrers or care providers) to visit, depending of their anticipated disutility from inpatient treatment. GPs choose their respective group, based on the degree of their altruistic motivation towards patients who experience disutility from inpatient care. The GPs who are more altruistically motivated will join the care provider group, while the extrinsically motivated GPs will belong to the group of referrers. This separating equilibrium is inefficient in the sense that too few patients visit the referrer group. A policy that prohibits kickbacks in this situation will decrease welfare, because the potential for cost savings through inpatient care would not be exploited. A subsidy for referrals will lower the referral price and induce more patients to consult a referrer. This policy would enable the efficient allocation to be achieved.

If patients are naïve or not capable of assessing the advantage of inpatient care over outpatient care, only a second-best allocation is feasible. Again, the equilibrium is separable, being split between two groups of referring GPs, each with a different degrees of altruism towards their patients' fates. Interestingly, in the second-best world with restricted patient capabilities, a prohibition against kickbacks can be an effective policy instrument, if complemented with adequate taxes. However, this need not necessarily be the case, since the combination of permitting kickbacks and setting appropriate taxes can also produce the second-best equilibrium. In particular, if a prohibition cannot be enforced, resorting to this instrument does not appear to be warranted. In the same vein, the scope for price regulation is limited. If prices are not correctly set, competition between providers will induce fee-splitting or side payments (*Pauly*, 1979). In such a scenario, price regulation would be undermined. It would therefore be a better policy to resort to taxes and subsidies.

The model could be extended to include a patient-specific probability of being severely ill. Depending on each individual's probability of illness, patients could select themselves into one of two patient groups: those who seek hospital care directly and those who visit their GP first. As long as it is assumed that the hospital cannot post a price for treating patients who have minor illnesses, the results of the present paper would apply. Competition in the outpatient sector and the hospital's profit maximization determine equilibrium prices and kickback payments. These factors inform patients about which provider they should choose. Patients who have a sufficiently high probability of becoming severely ill would prefer to seek inpatient care directly. The results change, however, if we assume that the hospital can post its own price for treating minor cases. In this case, the hospital is in direct competition with GPs and does not necessarily need the kickback channel to recruit patients with minor illnesses. In the separating equilibria described above, we would expect that referrers would be crowded out by the hospital, while care-providing GPs would remain in the outpatient market.

An equilibrium where heterogeneous GPs split into two separate groups according to their functional roles, by either treating their patients or referring them to hospitals with a diagnosis, is similar to some contractual arrangements in certain health care systems. In a market where providers

are allowed to engage in selective contracting and clients can enroll in specific health insurance plans, a self-sorting of providers and clients to different contracts might occur. As third-party stakeholders, insurers can design the reimbursement scheme for the providers and set appropriate user prices. Whether competition between insurers allows for prices that deviate from marginal costs is another story, which we leave for future research.

6. Appendix: Proof of Proposition 3

(i): Assume that all intrinsically motivated GPs are crowded out. At competitive prices, for $\alpha = 0$, $\kappa \geq p_H + s - c_G$ is required for patients to prefer referral to treatment by a GP. Furthermore, the patient's utility, if he is referred, is positive provided that $\kappa \geq \left(p_H - r + s + d \right) / \mu$. The maximum price the hospital can charge is $p_H = r - s$. At this price, we obtain $\kappa \geq r - \bar{c}_G$, $\kappa \geq 0$ and $p_r^R = d - \mu\kappa \leq 0$. This implies that no GP with $\alpha > 0$ will enter the market. Since all patients are referred and $s < \Delta c_G$, the equilibrium is efficient.

(ii): Assume that a separating equilibrium exists. Price competition between GPs implies $\pi_{GP} = 0$ for both groups: $\pi_{GP}^R = p_r^R - d + \mu\left(\kappa - \alpha s\right) = 0$ or $p_r^R = d - \mu\left(\kappa - \alpha s\right)$ and $\pi_{GP}^C = -d + \mu\left(p_t^C - \bar{c}_G \right) = 0$ or $p_t^C = d/\mu + \bar{c}_G$. Given the hospital price p_H, the utility of the referred patients is $u = r - s - p_H - p_r^R$, while the utility of the patients with minor illnesses who are treated by their GP is $u = (1-\mu)\left(r - s - p_H\right) + \mu\left(r - p_t^C\right)$. Additionally, the larger utility must be positive. A referral will increase patient utility at competitive prices, provided that $\kappa \geq p_H + (1+\alpha)s - c_G$. The hospital's profit is $\pi_H = (1-\mu)\left(p_H - c_H \right) + \mu F\left(\tilde{\alpha}\right)\left(p_H - \underline{c}_G - \kappa \right)$. It will pay kickbacks as long as $p_H - \underline{c}_G - \kappa \geq 0$. We obtain $0 \leq p_H - \underline{c}_G - \kappa \leq p_H - \underline{c}_G + \bar{c}_G - p_H - (1+\alpha)s = \Delta c_G - (1+\alpha)s$, which, for $s > \Delta c_G$, contradicts the assumption.

(iii): Extrinsically motivated GPs imitate $p_r^C = 0, p_t^C > 0$. This leads to $\pi_{GP}^C = -d + \mu\left(d/\mu + \bar{c}_G - \bar{c}_G \right) = 0$. For the marginal GP who does not provide care, we have $p_t^C - \bar{c}_G - d = \kappa - d - \tilde{\alpha}s$ or $\tilde{\alpha} = \left(\kappa + d/\mu\right)/s$. The profit maximizing hospital solves $\max_{\kappa} \left(p_H - \kappa - c_H^G \right)\mu x$, with $x = \tilde{\alpha}$, and the optimal kickback is $\kappa^* = \left(p_H - \underline{c}_G + d \right)/2$. For the pivotal GPs, we obtain $\tilde{\alpha} = \left(p_H - \underline{c}_G - d \right)/2s$. ■

7. References

Brekke, K.R., R. Nuscheler and O.R. Straume (2007), Gatekeeping in Health Care. *Journal of Health Economics* 26, 149-170.

Bussmann, K.-D. (2012), Unzulässige Zusammenarbeit im Gesundheitswesen durch „Zuweisung gegen Entgelt". Ergebnisse einer empirischen Studie im Auftrag des GKV-Spitzenverbandes, GKV-Spitzenverband (ed.), Berlin.

Chalkley, M. and J. Malcomson (1998), Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics* 17, 1–19.

Dulleck, U. and R. Kerschbamer (2006), On doctors, mechanics, and computer specialists: The economics of credence goods. Journal of Economic Literature, 5-42.

Evans, R.G. (1974), Supplier-induced demand: Some empirical evidence and implications. *The Economics of Health and Medical Care* 6, 162-173.

Felder, S. (2016), Kickbacks in Medical Expert Markets, Working Paper, University of Basel.

Garcia Mariñoso, B. and I. Jelovac (2003), GPs' payment contracts and their referral practice. *Journal of Health Economics* 22, 617–635.

Inderst, R. and M. Ottaviani (2012), Competition through Commissions and Kickbacks. *American Economic Review* 102(2): 780-809.

Malcomson, J.M. (2004), Health service gatekeepers. *RAND Journal of Economics* 35, 401–421.

McGuire, Th. (2000), Physician agency. In: Culyer, A. and Newhouse, J. (ed.). *Handbook of health economics*, 1st edition, 461-536.

Ng, Y.-K. (1983), Some Broader Issues of Social Choice. In Pattanaik, P.K. and M. Salles (Eds.), Social Choice and Welfare, 151-174. Amsterdam: North Holland.

Owen, B. M. (1977), Kickbacks, Specialization, Price Fixing, and Efficiency in Residential Real Estate Markets. *Stanford Law Review* 29(5): 931-967.

Pauly, M. (1979), The ethics and economics of kickbacks and fee splitting. *The Bell Journal of Economics* 10(1), 344-352.