Contracts for primary and secondary care physicians and equity-efficiency trade-offs*

Oddvar Kaarboe† Luigi Siciliani‡

December 3, 2021

Abstract

We analyse how payment systems for GPs (primary care physicians) and hospital specialists (secondary care) affect patients' inequalities in healthcare treatments, referrals and health. We present a model of contracting between a purchaser and two providers, a GP and a hospital specialist, with patients differing in severity and socioeconomic status. We assume that patients have high or low severity, and the GP only receives an informative signal on the severity of the patient following an examination. We investigate four health system configurations depending on whether the GP refers high-severity patients or high- and low-severity patients, and whether the specialist treats only high-severity patients or patients with any severity. We characterize possible equity-efficiency trade-offs arising from two policy interventions. We show that a tightening of a GP referral system generally increases allocative efficiency but also increases health inequities. A tightening of access to specialist services increases allocative efficiency and health inequities when the GP refers only high-severity patients, but has no effect on health inequities when the GP refers all patients.

Keywords: primary care; secondary care; equity; payment system; allocative efficiency.

JEL: I11, I14, I18.

*The paper is partly funded by the Research Council of Norway (288592).

†IGS and Department of Economics, University of Bergen, and HELED, University of Oslo, Norway. E-mail: oddvar.kaarboe@uib.no

‡Department of Economics and Related Studies, University of York, Heslington, York, UK. E-mail: luigi.siciliani@york.ac.uk.

1 Introduction

Reductions in health and healthcare inequalities are ubiquitous policy objectives. Despite these objectives, inequalities in healthcare utilization persist. For specialist visits the empirical evidence suggests a pro-rich gradient in most OECD countries (Van Doorslaer et al., 2004; Van Doorslaer and Masseria, 2004; Bago d'Uva and Jones, 2009; Devaux, 2015). For primary care visits, the results are more mixed, with some evidence suggesting pro-poor inequalities in a sub-set of countries (Van Doorslaer and Masseria, 2004; Bago d'Uva et al., 2009).

The question we ask in this study is how different payment systems for general practitioners (GPs) and hospital specialists affect inequalities in primary care and specialist visits, and the allocative efficiency of health systems. Different payment systems in primary care affect GP incentives to treat or refer patients to the specialist. Similarly, payment systems for specialists affect their incentives to treat the patient, or eventually refer the patient back to the GP. In turn, different combinations of payment systems in primary and secondary care generate different degrees of inequalities in treatments and referrals that translate into health inequalities, and different levels of welfare and therefore degree of allocative efficiency. More broadly, we investigate whether equity-efficiency trade-offs arise when changing health system configuration.

To answer our research question we present a model where a purchaser has contracts with two providers of health services, a GP and a hospital specialist. We assume that patients differ in severity, which can be high or low, and in socioeconomic status, which can also be high or low, giving four groups of patients. Patients cannot observe severity directly, and visit a GP when ill. The GP receives an informative signal on the severity of the patient following an examination. Critically, we assume that the signal the GP observes is more informative for patients with higher socioeconomic status, because these patients are better able to communicate the disease symptoms to the GP.[1]

[1] The systematic review by Deveugele et al. (2005) investigates the relationship between patients' socioeconomic status and the doctor-patient communication. It finds that GPs' communicative style is influenced by the way patients communicate: patients with higher socioeconomic status communicate more actively and show more affective expressiveness, eliciting more information from their doctor. The effects of good communication are studied in Greenfield et al. (1988). In their study, patients with diabetes were randomized to a previsit coaching session. In the coaching session, a clinical assistant reviewed the medical record with the patient and encouraged them to use the information gained to negotiate medical decisions with their doctor. Compared with non-coached, disease matched controls, intervention patients reported significantly fewer function limitations and lower hemoglobin HbA1 levels 6-12 weeks after the visit. Substantial improvements from baseline functioning were also observed in reported days lost from work among patients in the intervention group.

Based on the signal, the GP decides if to refer to a specialist or to treat the patient. For each patient referred by the GP, the specialist decides whether to treat or refer back based on their severity, which the specialist can observe perfectly. These assumptions give rise to four possible health system configurations: i) the GP refers only patients with high-severity signal and the specialist treats only high-severity patients; ii) the GP refers all patients, but the specialist treats only high-severity patients; iii) the GP refers only patients with high-severity signal, and the specialist treats all patients; iv) the GP refers all patients, and the specialist treats all patients.

We consider the most common payment systems that are in use. The GP is paid either by fee-for-service (FFS), capitation or a combination of the two. The hospital specialist is financed through a DRG-based payment system for the hospital specialist. Both the GP and the hospital specialist are altruistic and obtain utility both from patients' benefits of treatments and income.[2] We assume that the GP treatment cost is independent of severity (e.g. drug treatment), but that the specialist treatment cost is increasing in severity. Finally, we assume that if GP treatment for low-severity patients is delayed (due to the GP referring the patient to the specialist, and the specialist referring the patient back to the GP), patient's utility is reduced.

Our key findings are as follows. We generally find that health inequities are higher in health systems with tighter referrals where the GP refers only high-severity patients and lower in systems where specialists have stronger incentives to treat patients. More precisely, health inequities are highest under scenario i) when the GP refers patients with high-severity signal and the specialist treats only high-severity patients. Inequalities are intermediate in scenario iii) the GP refers patients with high-severity signal and the specialist treats all patients. There are no health inequities under scenario ii) or iv) when the specialist treats all patients regardless of whether the GP refers only high-severity patients or all patients.

In terms of welfare (allocative efficiency), we show that under minimal conditions welfare is highest when the GP referrals are tight and the specialist only treats the high-severity patients. Welfare is instead lowest when the referral system is loose so that the GP refers all patients and the specialist has incentives to treat all patients.

We then characterise policies that relate to tightening the referral system, or tightening the

[2]The idea that health care providers care (at least partially) about patients' utility or benefits of treatments has a long tradition in the economics literature on health care supply (Ellis and McGuire, 1986; Chalkley and Malcomson, 1998; Glazer, 2004; Kaarboe and Siciliani, 2011, Brekke et al., 2011).

access to specialist services, and possible equity-efficiency trade-offs that may arise as a result of implementing such policies. These policies are regularly discussed as interventions to contain costs and improve the sustainability of health spending. This is even more the case following the COVID-19 pandemic due to tightening of government budgets.

Consider a health system with a weak GP referral system where the GP refers all patients and specialists treat only high-severity patients, which corresponds to scenario ii). Then, inducing the GP to refer only patients with high-severity signal, which corresponds to a tightening of the referral system, implies a move from scenario ii) to scenario i). The introduction of a tighter referral system increases allocative efficiency but also increases health inequities, generating an equity-efficiency trade-off.

Similarly, consider a health system where the GP referral system is already tightened, but specialists have an incentive to treat all patients, which corresponds to scenario iii). Then, inducing the specialist to treat only high-severity patients, i.e. a tightening of access to specialist services, implies a move from scenario iii) to i), which increases allocative efficiency but also increases health inequities. Again, an equity-efficiency trade-off arises.

Finally, consider a health system with a loose GP referral system, and specialists have incentives to treat all patients, which is described under scenario iv). Then, tightening the GP referral system, a move from scenario iv) to iii), increases allocative efficiency but also increases health inequalities, which generates again an equity-efficiency trade-off. Instead, a tightening of access to specialist services, a move from scenario iv) to ii), will increase allocative efficiency, but has no effect on health inequities.[3]

In summary, an equity-efficiency trade-off is likely to arise in several circumstances. Our analysis is positive rather than normative. Rather than deriving an optimal payment system which would induce the implementation of the welfare maximising solution, we instead investigate the effects of realistic policy interventions, emphasising welfare and equity implications that may arise as a result.

The rest of the study is organised as follow. In Section 2, we provide a brief overview of the literature. In Section 3, we describe the key assumptions of the model. In Section 4, we

[3]It is unlikely that policymakers would move from scenario ii) to iii) or from iii) to ii). The former would involve tightening the referral system of the GP and at the same time easing access to specialist services. We therefore do not discuss these cases.

investigate provider incentives when the specialist treats only high-severity patients, while in Section 5 when the specialist treats all patients. Section 6 is devoted to the welfare analysis, and Section 7 concludes.

2 Related literature

Our study relates to different strands of the literature. Several studies have investigated the effect of different payment systems or the optimal payment system for health care providers, when doctors cannot observe severity directly but through an informative signal following an examination. Allard et al. (2011) compare the incentive properties of common payment systems for GPs. They find that capitation induces most referrals to expensive specialty care, and that fundholding induces almost as much referrals as capitation when the expected costs of primary care are high relative to secondary care. Mariñoso and Jelovac (2003), Malcomson (2004) and González (2010) also focus on the nature of GP's role in diagnosing patients and deciding whether to treat or refer. These studies derive optimal payment systems that simultaneously induce GPs to exert diagnosis effort and give incentives for efficient referral or treatment decisions, and discuss whether a gatekeeping system dominates free access to secondary care. Griebenow and Kifmann (2021) investigate the referral processes between a gatekeeping primary-care physician and a specialist when diagnostic signals are private information of the physicians. They show that welfare maximising optimal contracts involve a markup either to the GP for treating patients without referral or to the specialist for referring patients back to the GP. Godager et al. (2015) study the effect of competition on gatekeeping physicians' incentive to refer patients to a specialist, and show that the effect is in principle indeterminate. On one hand, competition induces the physician to refer more often to improve patient satisfaction, on the other hand they tend to earn more by treating patients themselves, weakening the incentive to refer. In their empirical analyses they show that the competition has negligible or small positive effects on total referrals. Brekke et al. (2007) study how gatekeeping affects hospital competition in the secondary care market. Patients, who are ex ante uninformed, can consult a GP to receive an (imperfect) diagnosis and obtain information about quality and specialization in the secondary care market. They show that hospital competition is amplified by higher GP attendance but

dampened by improved diagnosing accuracy. None of these studies investigate health inequalities and potential equity-efficiency trade-offs of different policy interventions, which is the focus of the current study.

Brekke et al. (2018) investigate the relationship between patients' socioeconomic status and GP provision of service. For patients in Norway with diabetes (type II) they show that patients with low education get shorter consultations but more medical tests, while patients with low income get less of both, and patients with low education/income get less services in monetary terms. Although mostly empirical, a theoretical framework is provided for patient-provider interaction where it is assumed that higher socioeconomic status increases the quality of the consultation. Chen and Lakdawalla (2019) investigate how altruism affects the way physicians respond to incentives and how patients' socioeconomic status mediates these responses. They show theoretically that patients' socioeconomic status systematically influences the way physicians respond to reimbursement changes. The model assumes that doctors care about the utility of the patient, which is a direct function of income, and therefore socioeconomic status. Using Medicare reimbursement changes they find that physicians facing an increase in reimbursement rates increase utilization more for richer relative to poorer patients.[4] We differ from these studies by using an informative signal framework, by allowing a more explicit interaction between the GP and the specialist, and by investigating the welfare implications of different policy interventions.

3 The Model

We present a model of provider behaviour with a GP and a hospital specialist serving a population of patients, which is normalized to one. Patients have high or low severity, $s \in \{\underline{s}, \overline{s}\}$, and high and low income[5], $i \in \{L, H\}$, giving four groups of patients. The proportion of patients with high and low severity with income i, is respectively equal to $\overline{\lambda}_i$ and $\underline{\lambda}_i$, with $\sum_{i=L,H}(\overline{\lambda}_i + \underline{\lambda}_i) = 1$.

We assume that there is a gatekeeping system and patients need to see a GP to access specialist care. This is common in many countries, like the Scandinavian countries, Canada,

[4] Since doctors do not generally have information on income within publicly funded systems we assume that doctors only care about patient health benefit.

[5] We use income as a proxy of socioeconomic status, therefore also including education, occupation etc.

Hungary, Netherlands, New Zealand, Poland, Portugal, Spain and United Kingdom. The GP who acts as gatekeeper decides whether to treat or refer a patient to the specialist. The specialist decides whether to treat the patient, or refer the patient back to the GP. The utility functions of the GP and the specialist are common knowledge. The GP and the specialist are paid by a health insurer and take the payment as given.

Timing. The timing of the game is as follows. First, the patient visits the GP. Second, the GP makes a decision about treatment or referral to the hospital specialist. Third, the specialist decides to treat the patient or to refer the patient back to the GP. If the patient is referred back, then the GP treats the patient.

All patients are ill and visit a GP. Patients do not know their severity. The GP does not observe patient's income but receives an informative signal on the severity of the patient, $\sigma \in \{\underline{s}, \overline{s}\}$. Define with $\Pr_i(\sigma|\, s)$ the probability of the doctor receiving a given signal σ conditional on a patient being of severity s and having income i. More precisely, the probability of the doctor receiving a *high*-severity signal conditional on the patient being *high* severity, for a given level of income i, is equal to $\Pr_i(\sigma = \overline{s}|\, s = \overline{s}) = \overline{\delta}_i > 0.5$. Similarly, the probability of the doctor receiving a *low*-severity signal conditional on the patient being *low* severity, for a given level of income i, is equal to $\Pr_i(\sigma = \underline{s}|\, s = \underline{s}) = \underline{\delta}_i > 0.5$. Therefore we assume that the signal is informative.[6]

We now state the probabilities across income groups. Consider a patient who has *high* severity. The probability of a doctor observing a *high*-severity patient and a *low*-severity patient is respectively equal to:

$$\Pr(\sigma = \overline{s}|\, s = \overline{s}) = \frac{\overline{\lambda}_L \overline{\delta}_L + \overline{\lambda}_H \overline{\delta}_H}{\overline{\lambda}_L + \overline{\lambda}_H}, \tag{1}$$

$$\Pr(\sigma = \underline{s}|\, s = \overline{s}) = \frac{\overline{\lambda}_L (1 - \overline{\delta}_L) + \overline{\lambda}_H (1 - \overline{\delta}_H)}{\overline{\lambda}_L + \overline{\lambda}_H}. \tag{2}$$

Instead, for a patient who has *low* severity, the probability of a doctor observing a *low*-severity

[6] Conversely, the probability of the doctor receiving a *low*-severity signal conditional on a patient being of *high* severity, for a given level of income i, is $\Pr_i(\sigma = \underline{s}|\, s = \overline{s}) = (1 - \overline{\delta}_i)$. The probability of the doctor receiving a *high*-severity signal conditional on the patient being of *low* severity, for a given level of income i, is equal to $\Pr_i(\sigma = \overline{s}|\, s = \underline{s}) = (1 - \underline{\delta}_i)$.

patient and a *high*-severity patient is respectively equal to:

$$\Pr(\sigma = \underline{s} | \, s = \underline{s}) = \frac{\underline{\lambda}_L \underline{\delta}_L + \underline{\lambda}_H \underline{\delta}_H}{\underline{\lambda}_L + \underline{\lambda}_H}, \tag{3}$$

$$\Pr(\sigma = \overline{s} | \, s = \underline{s}) = \frac{\underline{\lambda}_L (1 - \underline{\delta}_L) + \underline{\lambda}_H (1 - \underline{\delta}_H)}{\underline{\lambda}_L + \underline{\lambda}_H}. \tag{4}$$

Suppose that the GP observes a patient with a severity signal σ. What is the probability of the patient having severity s? Using Bayes' rule,[7] the probability of the GP facing a patient with severity s given the observed signal σ is equal to:

$$\Pr(s = \overline{s} | \, \sigma = \overline{s}) = \frac{\overline{\lambda}_L \overline{\delta}_L + \overline{\lambda}_H \overline{\delta}_H}{\overline{\lambda}_L \overline{\delta}_L + \overline{\lambda}_H \overline{\delta}_H + \underline{\lambda}_L (1 - \underline{\delta}_L) + \underline{\lambda}_H (1 - \underline{\delta}_H)},$$

$$\Pr(s = \underline{s} | \, \sigma = \underline{s}) = \frac{\underline{\lambda}_L \underline{\delta}_L + \underline{\lambda}_H \underline{\delta}_H}{\underline{\lambda}_L \underline{\delta}_L + \underline{\lambda}_H \underline{\delta}_H + \overline{\lambda}_L (1 - \overline{\delta}_L) + \overline{\lambda}_H (1 - \overline{\delta}_H)},$$

$$\Pr(s = \overline{s} | \, \sigma = \underline{s}) = \frac{\overline{\lambda}_L (1 - \overline{\delta}_L) + \overline{\lambda}_H (1 - \overline{\delta}_H)}{\overline{\lambda}_L (1 - \overline{\delta}_L) + \overline{\lambda}_H (1 - \overline{\delta}_H) + \underline{\lambda}_L \underline{\delta}_L + \underline{\lambda}_H \underline{\delta}_H},$$

$$\Pr(s = \underline{s} | \, \sigma = \overline{s}) = \frac{\underline{\lambda}_L (1 - \underline{\delta}_L) + \underline{\lambda}_H (1 - \underline{\delta}_H)}{\underline{\lambda}_L (1 - \underline{\delta}_L) + \underline{\lambda}_H (1 - \underline{\delta}_H) + \overline{\lambda}_L \overline{\delta}_L + \overline{\lambda}_H \overline{\delta}_H}.$$

We assume that $\overline{\delta}_H > \overline{\delta}_L$ and $\underline{\delta}_H > \underline{\delta}_L$, which implies that, for given severity, the signal is more informative for patients with high income because patients and doctors communicate better, which facilitates the assessment of the health state of the patient. Moreover, we assume that $\overline{\delta}_i > \underline{\delta}_i$, which implies that, for a given income, the signal is more informative for high severity patients than for low severity patients. This assumption is plausible. If the patient is in need of urgent care, the symptoms, such as pain level, fever, and unintended weight loss, are more likely to be detected by the doctor.

Patients' health benefit. The benefit for high-severity patients from being treated by a specialist and a GP is respectively equal to $B(\overline{s})$ and $b(\overline{s})$. We assume that specialists are better at treating high-severity patients, and $B(\overline{s}) > b(\overline{s})$. Similarly, the benefit for low-severity patients from being treated by a specialist and a GP is respectively equal to $B(\underline{s})$ and $b(\underline{s})$. Again, we assume that specialists are (weakly) better at treating low-severity patients, $B(\underline{s}) \geq b(\underline{s})$, but critically high-severity patients benefit more from being treated by a specialist, $B(\overline{s}) - b(\overline{s}) > B(\underline{s}) - b(\underline{s})$.

[7] $\Pr(s = s | \, \sigma = s) = \frac{\Pr(\sigma = s | s = s) \Pr(s)}{\Pr(\sigma = s | s = \underline{s}) \Pr(\underline{s}) + \Pr(\sigma = s | s = \overline{s}) \Pr(\overline{s})}$ where $\Pr(\overline{s}) = \overline{\lambda}_L + \overline{\lambda}_H$, $\Pr(\underline{s}) = \underline{\lambda}_L + \underline{\lambda}_H$.

Providers' cost. We assume that GP treatment cost is c, which is independent of severity (e.g. drug treatment), and that specialists treatment cost is equal to $C(s)$, which is increasing with severity and is more expensive than GP treatment, $C(\overline{s}) > C(\underline{s}) > c$.

Specialist utility function. We assume that specialists can always diagnose patient severity with no mistakes.[8] After diagnosis, the hospital specialist has two choices, either to treat or to refer the patient back to the GP. If the specialist treats a patient with severity s, her utility, defined with $V(\cdot)$, is given by

$$V(treat, s) = \begin{cases} T + P(s) - C(s) + \alpha^h B(s) & if \quad s = \overline{s} \\ T + P(s) - C(s) + \alpha^h B(s) - \Omega & if \quad s = \underline{s} \end{cases} \tag{5}$$

where $P(s)$ is a DRG tariff (or outpatient tariff), with $P(\overline{s}) \geq P(\underline{s}) \geq 0$, and $\alpha^h > 0$ is the specialist's degree of altruism (in line with previous literature, see Introduction for references). We assume that specialists have a disutility $\Omega \geq 0$ from treating a low-severity patient. For example, hospitals may have prioritisation protocols which give priority to high- rather than low-severity patients, and in many instances the latter can be treated in a primary care setting. Therefore, a specialist may feel guilty of treating a patient that could be treated in a less expensive setting. The disutility is likely to be higher in health systems with tight capacity constraints (as in some National Health Services), which implies that treating a low-severity patient may come at the cost of not treating a more severe patient. Instead, the disutility is likely to be low or zero in health systems with excess capacity. In each scenario, regardless of the patient severity or the decision to treat or refer, the specialist receives a non-negative fixed payment, $T \geq 0$ (e.g. a salary or a fixed budget).

If the specialist refers the patient back, her utility is:

$$V(referback, s) = T + \alpha^h \omega b(s), \quad \text{with } s \in \{\underline{s}, \overline{s}\}, \tag{6}$$

where ω is a weight related to the reduced utility due to delay in treatment, $0 < \omega < 1$. Lower values of ω imply larger losses of patient utility due to delayed benefits.

[8] In practice, specialists may also do some mistakes. Assuming that the specialist makes fewer mistakes than the GP would make the model more complicated but would not alter the key insights which are driven by the difference in the informativeness of the signal between the specialist and the GP.

The difference in specialist utility between treating the patient and referring the patient back to the GP is:

$$\begin{aligned}\Delta V(\overline{s}) &= V(treat,\overline{s}) - V(referback,\overline{s}) = P(\overline{s}) - C(\overline{s}) + \alpha^h \left(B(\overline{s}) - \omega b(\overline{s})\right), \\ \Delta V(\underline{s}) &= V(treat,\underline{s}) - V(referback,\underline{s}) = P(\underline{s}) - C(\underline{s}) + \alpha^h \left(B(\underline{s}) - \omega b(\underline{s})\right) - \Omega\end{aligned} \tag{7}$$

There are four possible scenarios. The specialist treats both severity types, only the high-severity type, only the low-severity type, or does not treat any patient at all. We rule the two latter (unlikely) scenarios by making the following assumptions.

A1 $\Delta V(\overline{s}) > 0$.

A2 $\Delta V(\overline{s}) > \Delta V(\underline{s})$.

Assumption A1 ensures that the specialist always has an incentive to treat a high-severity patient rather than referring the patient back to the GP: $\Delta V(\overline{s}) > 0$, or more extensively,

$$\alpha^h B(\overline{s}) + P(\overline{s}) > C(\overline{s}) + \alpha^h \omega b(\overline{s}). \tag{8}$$

The sum of the non-monetary patient benefits and the monetary ones, given by the DRG price, is larger than the treatment cost and the non-monetary cost for the patient from delayed GP treatment.

Assumption A2 ensures that the difference in specialist utility between treating and referring the patient back to the GP is higher for high-severity patients: $\Delta V(\overline{s}) > \Delta V(\underline{s})$, or more extensively:

$$\alpha^h \left(B(\overline{s}) - B(\underline{s})\right) + \Omega + P(\overline{s}) - P(\underline{s}) > C(\overline{s}) - C(\underline{s}) + \alpha^h \omega \left(b(\overline{s}) - b(\underline{s})\right). \tag{9}$$

The specialist benefits more from treating a high-severity patient compared to treating a low-severity patient if the differences in patient benefits weighted by altruism (including avoiding the disutility from treating a low-severity patient) and differences in monetary benefits, given by the difference in DRG tariffs, are larger than the difference in monetary costs of provision and non-monetary benefits from delayed treatment.

GP utility function. The utility of the GP, defined with $U(\cdot)$, from treating a patient with severity $s \in \{\underline{s}, \overline{s}\}$ is given by

$$U(treat, s) = t + p - c + \alpha^{gp} b(s), \tag{10}$$

where $p \geq 0$ is a fee received by the GP for each patient visit, and $t \geq 0$ is a fixed capitation payment. Instead, the utility of the GP from referring a patient to the specialist is

$$U(refer, s) = \begin{cases} t + \alpha^{gp} B(s) & if \quad s = \overline{s} \\ \alpha^{gp} \omega b(s) + t + p - c - k & if \quad s = \underline{s} \end{cases} \tag{11}$$

where $k \gtreqless 0$ captures a potential financial penalty for (inappropriately) referring a low-severity patient to the specialist.[9]

The rest of the analysis focuses on two plausible scenarios regarding specialist behaviour. In the first scenario, the specialist always has an incentive to treat high-severity patients and refer low-severity patients back to the GP, i.e. $\Delta V(\underline{s}) < 0$. In the second scenario, the specialist has an incentive to treat all referred patients, $\Delta V(\underline{s}) > 0$.

We discuss these two scenarios in turn respectively in Sections 3 and 4. For each scenario on the specialist behaviour, we distinguish two further sub-cases regarding the GP behaviour, whether the GP refers only high-severity patients or all patients the specialist. This gives four scenarios, which are also summarised in Figure 1:

1. the GP refers only high-severity patients and treats low-severity patients, and the specialist treats high-severity patients and refers low-severity patients back to the GP (scenario 1);

2. the GP refers both high- and low-severity patients, and the specialist treats high-severity patients and refers low-severity patients back to the GP (scenario 2);

3. the GP refers only high-severity patients and treats low-severity patients, and the specialist treats patients with high- and low-severity (scenario 3);

4. the GP refers both high- and low-severity patients, and the specialist treats patients with

[9] In health systems where there are no penalties, then $k = 0$, but in other systems k could be negative for example if the GP is paid for another visit when the specialist refers the patient back to the GP, $k = -p$.

high- and low-severity (scenario 4).

[Figure 1 here]

4 The specialist treats only high-severity patients

In this section we investigate scenarios 1 and 2 and assume that the specialist treats only high-severity patients $\Delta V(\underline{s}) < 0$, or more extensively:

$$\Delta V(\underline{s}) = P(\underline{s}) - C(\underline{s}) + \alpha^h \left(B(\underline{s}) - \omega b(\underline{s})\right) < \Omega. \tag{12}$$

This condition holds when the DRG price for low-severity patient is sufficiently low relative to the treatment cost and/or the disutility from treating a low severity patient is sufficiently high. In some health systems, such as in Norway or England, mixed or blended payment systems are in place for hospitals, where the DRG tariff covers only a proportion of the costs (e.g. 30-60%). In other systems, there may be penalties when hospitals admit high volume of patients, with the DRG price reducing to lower levels when volumes are above certain thresholds, which implies that the marginal tariff is lower. Even in health systems where the DRG tariff is set to cover the average cost, the presence of capacity constraints implies that there are protocols in place to prioritise hospital care for high-severity patients, which in turn implies that there is a (non-monetary) cost from admitting a low-severity patient.

The GP has to decide whether to treat or to refer to the specialist. The GP maximizes the expected utility where the expectation is taken over patient severity. If the GP *refers* the patient the expected utility for a given signal $\sigma \in \{\underline{s}, \overline{s}\}$ is equal to:

$$\mathbf{E}U(refer, \sigma) = t + \alpha^{gp} B(\overline{s}) \Pr(s = \overline{s} | \sigma) + (\alpha^{gp} \omega b(\underline{s}) + p - c - k) \Pr(s = \underline{s} | \sigma). \tag{13}$$

Instead, if the GP *treats* the patient, then the expected utility for a given signal σ is equal to:

$$\mathbf{E}U(treat, \sigma) = t + p - c + \alpha^{gp} b(\overline{s}) \Pr(s = \overline{s} | \sigma) + \alpha^{gp} b(\underline{s}) \Pr(s = \underline{s} | \sigma). \tag{14}$$

Define $\Delta \mathbf{E}U(\sigma) := \mathbf{E}U(refer, \sigma) - \mathbf{E}U(treat, \sigma)$ as the GP's expected utility gain or loss from

referring versus treating, for a given signal. Therefore, the GP refers the patient when

$$\begin{aligned}\Delta\mathbf{E}U(\sigma) &= \alpha^{gp}\left(B(\overline{s})-b(\overline{s})\right)\Pr(s=\overline{s}|\,\sigma) \\ &\quad -(\alpha^{gp}b(\underline{s})(1-\omega)+k)\Pr(s=\underline{s}|\,\sigma) \\ &\quad -(p-c)\left(1-\Pr(s=\underline{s}|\,\sigma)\right)\end{aligned} \tag{15}$$

is positive.

If the GP refers the patient, then the high-severity patient benefits more from the specialist treatment (first term). All low-severity patients that are referred to the specialist will be sent back to the GP and will suffer a utility loss due to delayed health benefit. The presence of penalties, $k > 0$, for referring low-severity patients further reduces GP's incentive to refer (second term). If GPs are paid by capitation, i.e. $t > 0$, $p = 0$, and $k = 0$, then the GP has always a financial incentive to refer the patient. If the GP is paid by FFS with a weakly positive price mark-up ($p \geq c$), then the GP has always a financial incentive to treat the patient (third term).

The GP refers the patient with a high-severity signal if $\Delta\mathbf{E}U(\sigma=\overline{s}) > 0$ and the GP refers the patient with a low-severity signal if $\Delta\mathbf{E}U(\sigma=\underline{s}) > 0$, that are respectively satisfied when

$$\begin{aligned}p &\leq \underline{p} := c+\frac{\alpha^{gp}\left(B(\overline{s})-b(\overline{s})\right)\Pr(s=\overline{s}|\,\sigma=\underline{s})-(\alpha^{gp}b(\underline{s})(1-\omega)+k)\Pr(s=\underline{s}|\,\sigma=\underline{s})}{(1-\Pr(s=\underline{s}|\,\sigma=\underline{s}))}, \\ p &\leq \overline{p} := c+\frac{\alpha^{gp}\left(B(\overline{s})-b(\overline{s})\right)\Pr(s=\overline{s}|\,\sigma=\overline{s})-(\alpha^{gp}b(\underline{s})(1-\omega)+k)\Pr(s=\underline{s}|\,\sigma=\overline{s})}{(1-\Pr(s=\underline{s}|\,\sigma=\overline{s}))}.\end{aligned}$$

Proposition 1 *Suppose that $\alpha^{gp}b(\underline{s})(1-\omega)+k > 0$, then $\overline{p} > \underline{p}$. If the fee received by the GP for each patient visit is low, i.e. $p \leq \underline{p}$, the GP always refers the patient to the specialist. If the fee is intermediate, i.e. $\underline{p} < p \leq \overline{p}$, the GP refers the patient to the specialist if she observes the high-severity signal, and she treats the patient if she observes the low-severity signal. If the fee is high, i.e. $p > \overline{p}$, the GP always treats the patient.*

See Appendix A1 for proof of Proposition 1. The move from a low to an intermediate fee could be interpreted as the introduction of a FFS system. The case with an intermediate GP fee for a visit corresponds to scenario 1 in Figure 1, and the case with a low GP fee for a visit

corresponds to scenario 2 in Figure 1.[10] We discuss these two scenarios in turn in the next two sub-sections 4.1 and 4.2.

Finally, notice that $\underline{p}$ could be negative if postponing treatment generates significant losses in patient benefits or if the financial penalties for referring low-severity patients are sufficiently high. In turn, this implies that the GP refers only the patient with the high-severity signal, even if the GP is paid only by capitation, and receives no fee for each patient visit.

4.1 GP refers only patients with high severity signal to the specialist

The total number of referrals R is given by the probability of a high-severity signal:[11]

$$R = \Pr(\sigma = \overline{s}) = \overline{\lambda}_L \overline{\delta}_L + \overline{\lambda}_H \overline{\delta}_H + \underline{\lambda}_L(1 - \underline{\delta}_L) + \underline{\lambda}_H(1 - \underline{\delta}_H). \tag{16}$$

The number of referrals for each income group is: $R_i = \overline{\lambda}_i \overline{\delta}_i + \underline{\lambda}_i(1 - \underline{\delta}_i)$, $i = L, H$. Define $\overline{\Lambda}_i := \frac{\overline{\lambda}_i}{\overline{\lambda}_i + \underline{\lambda}_i}$ and $\underline{\Lambda}_i := \frac{\underline{\lambda}_i}{\overline{\lambda}_i + \underline{\lambda}_i}$ as the incidence of high- and low-severity in income group $i = L, H$. The proportion of GP referrals within each income group, defined with r_i, is then $r_i = \frac{R_i}{\overline{\lambda}_i + \underline{\lambda}_i} = \overline{\Lambda}_i \overline{\delta}_i + \underline{\Lambda}_i(1 - \underline{\delta}_i)$, $i = L, H$. Since low-severity patients are sent back to the GP, the proportion of specialist treatment within each income group, defined with v_i, is given by $v_i = \overline{\Lambda}_i \overline{\delta}_i$, $i = L, H$. The proportion of GP treatment within each income group is therefore $g_i = 1 - v_i = 1 - \overline{\Lambda}_i \overline{\delta}_i$, $i = L, H$. Using the above, the income-related *inequalities* in the GP's referrals rates are:

$$r_H - r_L = \left(\overline{\delta}_H - \overline{\delta}_L\right) \overline{\Lambda}_H - (\underline{\delta}_H - \underline{\delta}_L)\underline{\Lambda}_H + \overline{\delta}_L \left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) - (1 - \underline{\delta}_L) \left(\underline{\Lambda}_L - \underline{\Lambda}_H\right). \tag{17}$$

Inequalities in GP referrals depend on the accuracy of the signal across the two income groups (given by the first and second terms) and the incidence of low and high severity in each income group (given by the third and fourth term).

The income-related inequalities in the proportion of specialist treatment is:

$$v_H - v_L = \overline{\Lambda}_H \overline{\delta}_H - \overline{\Lambda}_L \overline{\delta}_L. \tag{18}$$

[10] Below, we do not discuss the scenario when the fee is high enough that the GP has an incentive to treat also the high-severity patients, since we do not consider it a plausible scenario.

[11] Given the population of patients is normalised to one, the total probability of a high-severity signal is $\Pr(\sigma = \overline{s}) = \Pr(\sigma = \overline{s} | s = \overline{s}) \Pr(s = \overline{s}) + \Pr(\sigma = \overline{s} | s = \underline{s}) \Pr(s = \underline{s})$, where $\Pr(s = \overline{s}) = \overline{\lambda}_L + \overline{\lambda}_H$ and $\Pr(s = \underline{s}) = \underline{\lambda}_L + \underline{\lambda}_H$.

Whether the proportion of specialist treatment is higher in the high-income group is also in principle indeterminate. For example, if the high-income group has a lower incidence of high severity, then the proportion of specialist treatment will be higher only if the accuracy effect dominates over the incidence effect. The income-related gradient in GP treatment is the reverse of the gradient in specialist treatment, $g_H - g_L = -(v_H - v_L)$.

We can decompose the income-related inequalities in specialist treatment in two components:

$$v_H - v_L = \overline{\Lambda}_H \left(\overline{\delta}_H - \overline{\delta}_L\right) + \left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) \overline{\delta}_L. \tag{19}$$

We refer to the first component as *inequities* in specialist treatment, and to the second term to inequalities as these reflect severity incidence. Within many publicly funded health systems, healthcare is supposed to be allocated on need, not ability to pay. Inequalities that arise due to a higher incidence of a disease reflect differences in need, and do not count as inequities. Instead, we refer to inequities for differences in specialist treatment that are due to patient ability to convey the high-severity signal. We therefore conclude that there are pro-rich inequities in specialist treatment and pro-poor inequities in GP treatment.[12]

The expected benefit from treatment for each income group is

$$\mathcal{B}_i = \overline{\Lambda}_i \left[\overline{\delta}_i B(\overline{s}) + (1 - \overline{\delta}_i) b(\overline{s})\right] + \underline{\Lambda}_i b(\underline{s}) \left[\underline{\delta}_i + (1 - \underline{\delta}_i)\, \omega\right], \; i = L, H, \tag{20}$$

which gives the benefit across high- and low-severity patients weighted by the severity incidences, and is increasing in the precision of the GP signal. The income-related inequalities in health

[12] Whether there is a pro-rich or pro-poor gradient in GP referrals is still indeterminate even if the $\overline{\Lambda}_H = \overline{\Lambda}_L = \overline{\Lambda}$. The gradient in referrals simplifies to: $r_H - r_L = \left(\overline{\delta}_H - \overline{\delta}_L\right) \overline{\Lambda} - (\underline{\delta}_H - \underline{\delta}_L)\underline{\Lambda}$. The gradient depends on the accuracy of the severity signal, weighted by the incidence of high- and low-severity patients, across the two income groups. The difference in referrals consists of two terms. The first term measures the precision of the high-severity signal of the high income group relative to the high-severity signal of the low income group. A more precise high-severity signal for the high income group, relative to the high-severity signal of the low income group, contributes towards a pro-rich gradient in specialist referrals. The second term measures the mistakes, namely the low-severity patients for whom the GP observes a high-severity signal in the two income groups and refers to the specialist. A less precise low-severity signal of the low income group increases the probability of mistakes and hence contributes towards a pro-poor gradient in specialist referrals. If the incidence of high- and low-severity is the same (i.e. $\overline{\Lambda} = \underline{\Lambda}$) then the gradient in referrals is pro-rich, given our assumption that the severity signal is more informative when the patient has high severity. This is also the case if the incidence of high severity is higher than the incidence of low severity. But a pro-poor gradient can arise if the incidence of low severity is sufficiently high relative to high severity.

benefits are given by (see Appendix A2):

$$\begin{aligned}\mathcal{B}_H - \mathcal{B}_L &= \overline{\Lambda}_H \left(\overline{\delta}_H - \overline{\delta}_L\right) \left[B(\overline{s}) - b(\overline{s})\right] + \underline{\Lambda}_H \left(\underline{\delta}_H - \underline{\delta}_L\right) b(\underline{s}) \left(1 - \omega\right) \\ &\quad - \left(\overline{\Lambda}_L - \overline{\Lambda}_H\right) \left[\overline{\delta}_L B(\overline{s}) + (1 - \overline{\delta}_L) b(\overline{s})\right] - \left(\underline{\Lambda}_L - \underline{\Lambda}_H\right) b(\underline{s}) \left[\underline{\delta}_L - (1 - \underline{\delta}_L)\omega\right].\end{aligned} \tag{21}$$

The first line relates to accuracy of the GP signals, and both terms are positive. The first term captures that patients with high severity are more likely to benefit from specialist treatment if they have high income. The second term is related to the fact that the GP receives a more precise signal of the patient being of low severity when s/he has high income. This implies that the GP refers less often a high income patient with low severity to the specialist. As a consequence, fewer low severity patients with high income experience delayed treatments. This contributes to pro-rich inequities in health benefits. The second line is due to differences in incidences, and therefore do not contribute to pro-rich inequities. We therefore conclude that there are pro-rich inequities in health benefits. We summarise in the following proposition (scenario 1 in Figure 1).

Proposition 2 *Let the GP fee for a visit be such that $\underline{p} < p \leq \overline{p}$ so that the GP refers only patients when a high-severity signal is observed, and the specialist only treats high-severity patients. Then, there are pro-rich inequities in specialist treatment, and in health benefit from treatment. There are pro-poor inequities in GP treatment.*

4.2 GP refers all patients to the specialist

In this case all patients are referred to the specialist, i.e. $R_H = \overline{\lambda}_H + \underline{\lambda}_H$, $R_L = \overline{\lambda}_L + \underline{\lambda}_L$, who will only treat high-severity patients. The proportion of referrals to the specialist for low and high-income patients are $r_H = 1$, $r_L = 1$, $r_H - r_L = 0$. Since low-severity patients are sent back to the GP, the proportion of specialist treatment in each income group is $v_i = \overline{\Lambda}_i$, $i = L, H$, and the gradient is

$$v_H - v_L = \overline{\Lambda}_H - \overline{\Lambda}_L. \tag{22}$$

Whether the proportion of specialist treatment is higher in the high-income group depends on severity incidence, and therefore such differences do not constitute a source of inequity.

The proportion of GP treatment in each income group is: $g_i = 1 - \overline{\Lambda}_i = \underline{\Lambda}_i$, $i = L, H$, and the gradient is

$$g_H - g_L = \underline{\Lambda}_L - \underline{\Lambda}_H. \tag{23}$$

The expected benefit from treatment for each income group is $\mathcal{B}_i = \overline{\Lambda}_i B(\overline{s}) + \underline{\Lambda}_i b(\underline{s})\omega$, $i = L, H$, and the gradient is

$$\mathcal{B}_H - \mathcal{B}_L = \left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) B(\overline{s}) + \left(\underline{\Lambda}_H - \underline{\Lambda}_L\right) b(\underline{s})\omega. \tag{24}$$

Notice that the difference in benefit is amplified by the delay ω. We summarise in the following proposition (scenario 2 in Figure 1).

Proposition 3 *Let the GP fee for a visit be sufficiently low, such that $p < \underline{p}$, so that the GP refers all patients, and the specialist only treats high-severity patients. Inequalities in treatment and benefit are related to differences in incidence of high severity across income groups, and there are therefore no inequities in treatment and health benefit.*

5 The specialist treats all patients

In this section we assume that $\Delta V(\underline{s}) > 0$, so that the specialist has an incentive to treat all referred patients. The GP has to decide whether to treat or refer the patient to the specialist. If the GP *refers* the patient the expected utility for a given signal $\sigma \in \{\underline{s}, \overline{s}\}$ is equal to:

$$\mathbf{E}U(refer, \sigma) = \alpha^{gp} \left[B(\overline{s}) \Pr(\overline{s}|\, \sigma) + B(\underline{s}) \Pr(\underline{s}|\, \sigma)\right]. \tag{25}$$

Instead, if the GP *treats* the patient, then the expected utility for a given signal σ is equal to:

$$\mathbf{E}U(treat, \sigma) = p - c + \alpha^{gp} b(\overline{s}) \Pr(\overline{s}|\, \sigma) + \alpha^{gp} b(\underline{s}) \Pr(\underline{s}|\, \sigma). \tag{26}$$

Therefore, the GP refers the patient when

$$\Delta\mathbf{E}U(\sigma) = \alpha^{gp} \left(B(\overline{s}) - b(\overline{s})\right) \Pr(\overline{s}|\, \sigma) + \alpha^{gp} \left(B(\underline{s}) - b(\underline{s})\right) \Pr(\underline{s}|\, \sigma) - p + c > 0. \tag{27}$$

More precisely, the GP refers the patient with a high-severity signal if $\Delta EU\,(\sigma = \overline{s}) > 0$ and the patient with a low-severity signal if $\Delta EU\,(\sigma = \underline{s}) > 0$. These are respectively satisfied when

$$\begin{aligned} p &< \widetilde{p} := c + \alpha^{gp}\left(B(\overline{s}) - b(\overline{s})\right)\Pr(\overline{s}|\,\overline{s}) + \alpha^{gp}\left(B(\underline{s}) - b(\underline{s})\right)\Pr(\underline{s}|\,\overline{s}), \\ p &< \underset{\sim}{p} := c + \alpha^{gp}\left(B(\overline{s}) - b(\overline{s})\right)\Pr(\overline{s}|\,\underline{s}) + \alpha^{gp}\left(B(\underline{s}) - b(\underline{s})\right)\Pr(\underline{s}|\,\underline{s}), \end{aligned}$$

The following proposition characterizes the GP referral and treatment decisions.

Proposition 4 $\widetilde{p} > \underset{\sim}{p} > 0$. *If the GP fee for a visit is low, i.e.* $0 \leq p \leq \underset{\sim}{p}$, *the GP always refers the patient to the specialist. If the GP fee is intermediate, i.e.* $\underset{\sim}{p} < p \leq \widetilde{p}$, *the GP refers the patient to the specialist if she observes the high-severity signal, and she treats the patient if she observes the low-severity signal. If the GP fee is high, i.e.* $p > \widetilde{p}$, *the GP always treats the patient.*

See Appendix A3 for proof of Proposition 3. The GP has always an incentive to refer under capitation, when $p = 0$, and this is the case under FFS if the fee is set equal to the marginal cost, $p = c$. This arises because patients benefit more from the specialist treatment than GP treatment, and there is no risk that the patient is referred back to the GP as by assumption the specialist treats all referred patients.[13] The case with an intermediate GP fee for a visit corresponds to scenario 3 in Figure 1, and the case with a low GP fee for a visit corresponds to scenario 4 in Figure 1. We discuss these two scenarios in turn in the next two sub-sections 5.1 and 5.2.

5.1 GP refers only patients with a high severity signal to the specialist

If the GP fee for a visit is intermediate, i.e. $\underset{\sim}{p} < p \leq \widetilde{p}$, the total number of referrals R is, as in Section 3.1, $R = \overline{\lambda}_L\overline{\delta}_L + \overline{\lambda}_H\overline{\delta}_H + \underline{\lambda}_L(1-\underline{\delta}_L) + \underline{\lambda}_H(1-\underline{\delta}_H)$, which again can be split across income groups, $R_i = \overline{\lambda}_i\overline{\delta}_i + \underline{\lambda}_i(1-\underline{\delta}_i)$, $i = L, H$. The proportion of GP referrals within each income group are equal to $r_i = \overline{\delta}_i\overline{\Lambda}_i + (1-\underline{\delta}_i)\underline{\Lambda}_i$, and the income-related inequalities in GP referrals are

[13] Similarly to Section 4, we do not discuss the scenario when the GP fee is high enough that the GP has an incentive to treat also the high-severity patients, since we do not consider it a plausible scenario.

equal to:

$$r_H - r_L = \left(\overline{\delta}_H - \overline{\delta}_L\right)\overline{\Lambda}_H - (\underline{\delta}_H - \underline{\delta}_L)\underline{\Lambda}_H + \overline{\delta}_L\left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) - (1 - \underline{\delta}_L)\left(\underline{\Lambda}_L - \underline{\Lambda}_H\right). \tag{28}$$

The income-related inequalities in GP referrals are identical to the scenario when the GP refers only high-severity patients and the specialist refers low-severity patients back to the GP (see Section 4.1, equation (17)), and therefore depends on the incidence of high severity in each income group and the accuracy of the signal across the two income groups, and is in principle indeterminate. Since low-severity patients are not sent back to the GP, any income-related inequality in GP referrals translates into inequalities in the proportion of specialist treatment, with $v_i = r_i$, and in the proportion of GP treatment, with $g_i = 1 - r_i = 1 - v_i$, so that

$$r_H - r_L = v_H - v_L = g_L - g_H. \tag{29}$$

We again decompose inequalities in specialist treatments between inequalities due to income (inequities) in the first two terms in (28) and inequalities due to differences in severity incidence in the last two terms in (28). Income-related inequities in treatment depend on the accuracy of the signal. Since the signal is more accurate for high-income patients, then $\left(\overline{\delta}_H - \overline{\delta}_L\right)\overline{\Lambda}_H > 0$ and $(\underline{\delta}_H - \underline{\delta}_L)\underline{\Lambda}_H > 0$: patients with high-income are more likely to visit a specialist if they have high severity but less likely to visit a specialist if they have low severity. If the incidence of low-severity patients is sufficiently low (high), this leads to pro-rich (pro-poor) inequities in specialist visits.

The expected benefit from treatment for each income group is

$$\mathcal{B}_i = \overline{\Lambda}_i\left[\overline{\delta}_i B(\overline{s}) + (1 - \overline{\delta}_i)b(\overline{s})\right] + \underline{\Lambda}_i\left[\underline{\delta}_i b(\underline{s}) + (1 - \underline{\delta}_i)B(\underline{s})\right], \; i = L, H, \tag{30}$$

and inequalities in health benefit are given by

$$\begin{aligned} \mathcal{B}_H - \mathcal{B}_L \;=\; & \overline{\Lambda}_H\left(\overline{\delta}_H - \overline{\delta}_L\right)\left[B(\overline{s}) - b(\overline{s})\right] - \underline{\Lambda}_H\left(\underline{\delta}_H - \underline{\delta}_L\right)\left[B(\underline{s}) - b(\underline{s})\right] \\ & - \left(\overline{\Lambda}_L - \overline{\Lambda}_H\right)\left[\overline{\delta}_L B(\overline{s}) + (1 - \overline{\delta}_L)b(\overline{s})\right] + \left(\underline{\Lambda}_H - \underline{\Lambda}_L\right)\left[\underline{\delta}_L b(\underline{s}) + (1 - \underline{\delta}_L)B(\underline{s})\right]. \end{aligned} \tag{31}$$

We can again decompose inequalities in health benefits between inequalities due to income (inequities) in the first line and inequalities due to differences in severity incidence in the second line. Health inequities depend on the accuracy of the signal. Since the signal is more accurate for high-income patients, the first term in the first line is positive and the second term is negative: patients with high severity are more likely to benefit from specialist treatment if they are of high income. However, low-severity patients are more likely to benefit from specialist treatment if they are of low income. This follows because their low-severity signal observed by the GP is less precise, so that more low-income patients are referred to the specialist. We summarise in the following proposition (scenario 3 in Figure 1).

The following proposition isolates the gradient due to the accuracy of the signal.

Proposition 5 *Let the GP fee for a visit be intermediate, $\underset{\sim}{p} < p \leq \widetilde{p}$, so that the GP refers only patients when a high-severity signal is observed, and the specialist treats patients with any severity. Then there are pro-rich (pro-poor) inequities in specialist treatment and health benefits if the incidence of low-severity patients is sufficiently low (high).*

Finally, notice that relative to scenario 1, income-related health inequities are always higher in scenario 1 than in the current scenario 3. This follows immediately by comparing (21) with (31), as the difference in the gradient is given by $\underline{\Lambda}_H \left(\underline{\delta}_H - \underline{\delta}_L\right) b(\underline{s})\omega + \underline{\Lambda}_H \left(\underline{\delta}_H - \underline{\delta}_L\right) \left[B(\underline{s}) - b(\underline{s})\right] > 0$.

5.2 GP refers all patients to the specialist

If the GP fee for a visit is low, $p < \underset{\sim}{p}$, all patients are referred to the specialist who will treat them, i.e. $R_H = \overline{\lambda}_H + \underline{\lambda}_H$, $R_L = \overline{\lambda}_L + \underline{\lambda}_L$. The proportion of referrals to the specialist for low and high-income is $r_H = r_L = 1$. The proportion of specialist treatment in the high- and low-income groups is also $v_H = v_L = 1$. Conversely, the proportion of GP treatment in the high- and low-income groups is $g_H = g_L = 0$. The expected benefit from treatment for high- and low-income groups is $\mathcal{B}_H = \overline{\Lambda}_H B(\overline{s}) + \underline{\Lambda}_H B(\underline{s})$, $\mathcal{B}_L = \overline{\Lambda}_L B(\overline{s}) + \underline{\Lambda}_L B(\underline{s})$ and inequalities in health benefits are given by

$$\mathcal{B}_H - \mathcal{B}_L = \left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) B(\overline{s}) + \left(\underline{\Lambda}_H - \underline{\Lambda}_L\right) B(\underline{s}). \tag{32}$$

Since all patients receive specialist treatment, the only gradient in benefits is due to differences in severity incidences. We summarise in the following proposition (scenario 4 in Figure 1).

Proposition 6 *If the GP fee for a visit is low, $p < \underset{\sim}{p}$, so that the GP refers all patients, and the specialist treats all patients, there are no inequities in GP referrals and specialist treatment. Inequalities in health benefits are driven by differences in severity incidence across income groups.*

In the next section, we discuss welfare implications and identify possible equity-efficiency trade-offs.

6 Welfare

We adopt a utilitarian welfare function which we define as the difference between patient benefit and provider costs. With no uncertainty about the severity of the patient, it is welfare improving for a patient to be treated by a specialist relative to GP treatment if the difference in net benefit, defined with

$$\Delta NB(s) = B(s) - C(s) - (b(s) - c), \tag{33}$$

is positive. In the following, we assume that:

A3 $\Delta NB(\overline{s}) > 0$.

A4 $\Delta NB(\underline{s}) < 0$.

Assumption A3 implies that the benefit from being treated by a specialist relative to a GP is positive for the high-severity patients, while assumption A4 implies that it is negative for low-severity patients. Under these assumptions, it is optimal from a utilitarian welfare perspective that the specialist treats the high-severity patients, and the GP treats the low-severity patients. We refer to this allocation of patients as the "first best".

The total welfare under the first best solution is given by:

$$W^{fb} = \left(\overline{\lambda}_H + \overline{\lambda}_L\right)\left[B(\overline{s}) - C(\overline{s})\right] + \left(\underline{\lambda}_H + \underline{\lambda}_L\right)\left[b(\underline{s}) - c\right]. \tag{34}$$

Using the welfare under the first best as a benchmark, we compare welfare under the four scenarios identified in Sections 4 and 5 against this benchmark. We define $W(s, s)$ as the

welfare where the first argument refers to the GP decision to *refer* a patient with given severity s, and the second argument refers to the specialist decision to *treat* a patient with given severity s. Hence, $W(\overline{s},\overline{s})$, $W(\overline{s},all)$, $W(all,\overline{s})$, $W(all,all)$ denote welfare when respectively i) the GP refers high-severity patients, and the specialist treats only high-severity patients; ii) the GP refers high-severity patients, and the specialist treats all patients; iii) the GP refers all patients, and the specialist treats only high-severity patients; iv) the GP refers all patients, and the specialist treats all patients. More explicitly, we obtain the following expressions:

$$\begin{aligned}
W(\overline{s},\overline{s}) &= \left(\overline{\lambda}_H\overline{\delta}_H+\overline{\lambda}_L\overline{\delta}_L\right)\left(B(\overline{s})-C(\overline{s})\right)+\left(\overline{\lambda}_H(1-\overline{\delta}_H)+\overline{\lambda}_L(1-\overline{\delta}_L)\right)\left(b(\overline{s})-c\right) \\
&\quad +\left(\underline{\lambda}_H+\underline{\lambda}_L\right)\left(b(\underline{s})-c\right)-\left[\underline{\lambda}_H(1-\underline{\delta}_H)+\underline{\lambda}_L(1-\underline{\delta}_L)\right]b(\underline{s})\left(1-\omega\right), \\
W(\overline{s},all) &= \left(\overline{\lambda}_H\overline{\delta}_H+\overline{\lambda}_L\overline{\delta}_L\right)\left(B(\overline{s})-C(\overline{s})\right)+\left(\overline{\lambda}_H(1-\overline{\delta}_H)+\overline{\lambda}_L(1-\overline{\delta}_L)\right)\left(b(\overline{s})-c\right) \\
&\quad +\left(\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L\right)\left(b(\underline{s})-c\right)+\left(\underline{\lambda}_H(1-\underline{\delta}_H)+\underline{\lambda}_L(1-\underline{\delta}_L)\right)\left(B(\underline{s})-C(\underline{s})\right), \\
W(all,\overline{s}) &= \left(\overline{\lambda}_H+\overline{\lambda}_L\right)\left[B(\overline{s})-C(\overline{s})\right]+\left(\underline{\lambda}_H+\underline{\lambda}_L\right)\left[b(\underline{s})(1-(1-\omega))-c\right], \\
W(all,all) &= \left(\overline{\lambda}_H+\overline{\lambda}_L\right)\left[B(\overline{s})-C(\overline{s})\right]+\left(\underline{\lambda}_H+\underline{\lambda}_L\right)\left[B(\underline{s})-C(\underline{s})\right].
\end{aligned} \tag{35}$$

In $W(all,\overline{s})$, we can write $b(\underline{s})(1-(1-\omega))-c=\omega b(\underline{s})-c$, as the benefit for low-severity patients is discounted as they are systematically sent back.

After computing $\Delta W(s,s)=W(s,s)-W^{fb}$, straightforward calculations give:

$$\begin{aligned}
\Delta W(\overline{s},\overline{s}) &= -\left(\overline{\lambda}_H\left(1-\overline{\delta}_H\right)+\overline{\lambda}_L\left(1-\overline{\delta}_L\right)\right)\Delta NB(\overline{s}) \\
&\quad -\left(\underline{\lambda}_H(1-\underline{\delta}_H)+\underline{\lambda}_L(1-\underline{\delta}_L)\right)b(\underline{s})\left(1-\omega\right), \\
\Delta W(\overline{s},all) &= -\left(\overline{\lambda}_H\left(1-\overline{\delta}_H\right)+\overline{\lambda}_L\left(1-\overline{\delta}_L\right)\right)\Delta NB(\overline{s}) \\
&\quad +\left(\underline{\lambda}_H(1-\underline{\delta}_H)+\underline{\lambda}_L(1-\underline{\delta}_L)\right)\Delta NB(\underline{s}), \\
\Delta W(all,\overline{s}) &= -\left(\underline{\lambda}_H+\underline{\lambda}_L\right)b(\underline{s})\left(1-\omega\right), \\
\Delta W(all,all) &= \left(\underline{\lambda}_H+\underline{\lambda}_L\right)\Delta NB(\underline{s}).
\end{aligned} \tag{36}$$

Let's consider the welfare loss when the GP refers all patients who are then treated by the specialist, i.e. $\Delta W(all,all)$. In this scenario, the total welfare loss depends on the number of low-severity patients treated by the specialist, multiplied by the welfare loss for each patient

from being treated by a specialist rather than the GP.

If all patients are referred, but the specialist only treats the high severity patients, $\Delta W(all, \overline{s})$, the welfare loss depends on the delay in treatment of the low-severity patients who are referred back to the GP, and is independent of the precision of the severity signals. If delay in treatment is costless, i.e. $\omega = 1$, there is no welfare loss.

If the GP only refers when a high-severity signal is observed and the specialist only treats high-severity patients, $\Delta W(\overline{s}, \overline{s})$, the welfare loss consists of two parts. The first part is the welfare loss that occurs since some high-severity patients are treated by the GP (who receive a signal that these patients are of low severity), while they should be treated by the specialist. The second part of the welfare loss depends on the number of low-severity patients who see their treatment delayed because they are referred to the specialist who sends them back to the GP.

Finally, if the GP only refers when a high-severity signal is observed but the specialist treats every referred patient, $\Delta W(\overline{s}, all)$, the welfare loss is related to the the GP's misinterpretation of the signals: A share of the low-severity patients is treated by the specialist, and a share of high-severity patients are treated by the GP. The less precise are the signals, the higher is the welfare loss. Moreover, the welfare loss increases with the difference in net benefits of being treated by the "wrong" doctor.

To characterize cases where an equity-efficiency trade-off arises, we collect earlier results on *inequities* in patient benefit across income groups. That is we disregard inequalities in benefits that are due to differences in severity incidences. Let $\Delta\mathcal{B}(s, s) := B_H - B_{L|\Lambda_H=\Lambda_L}$, i.e. the income-related *inequity* in health benefits, where the first argument refers to the GP's decision to refer a patient with given severity s, and the second argument refers to the specialist decision to treat a patient with given severity s. From equations (21), (24), (31) and (32) we obtain:

$$\begin{aligned}
\Delta\mathcal{B}(\overline{s}, \overline{s}) &= \overline{\Lambda}_H \left(\overline{\delta}_H - \overline{\delta}_L\right) [B(\overline{s}) - b(\overline{s})] + \underline{\Lambda}_H \left(\underline{\delta}_H - \underline{\delta}_L\right) b(\underline{s}) \left(1 - \omega\right) > 0, \qquad (37) \\
\Delta\mathcal{B}(\overline{s}, all) &= \overline{\Lambda}_H \left(\overline{\delta}_H - \overline{\delta}_L\right) [B(\overline{s}) - b(\overline{s})] - \underline{\Lambda}_H \left(\underline{\delta}_H - \underline{\delta}_L\right) [B(\underline{s}) - b(\underline{s})] \gtrless 0, \\
\Delta\mathcal{B}(all, all) &= \Delta\mathcal{B}(all, \overline{s}) = 0.
\end{aligned}$$

The health gradients are the direct result of inequalities in specialist treatments in the four scenarios.[14] Suppose first that the GP refers only the high-severity patients. Then, if the specialist treats only the high-severity patients, there is a pro-rich gradient in health benefits. If instead the specialist treats all patients, then the gradient can be either pro-rich or pro-poor. Since the signal is more accurate for high-income patients, these patients are more likely to visit a specialist if they have high severity but less likely to visit a specialist if they have low severity. The sign of the gradient does however also depend on the incidence of low and high severity. More specifically, if the incidence of low-severity patients is sufficiently high (low), this leads to pro-poor (pro-rich) inequities in specialist visits. Finally, if the GP refers patients with high and low severity, then there is no health gradient. We can also show that income-related health inequities are highest when the GP refers only patients with high severity and the specialist also treats patients only with high severity, i.e. $\Delta\mathcal{B}(\overline{s},\overline{s}) > \Delta\mathcal{B}(\overline{s},all)$.[15]

We consider the introduction of two policies. The first relates to tightening the access to specialist services, and the second to tightening the referral system. We discuss these in turn in Propositions 7 and 8.

Proposition 7 *Consider a policy that tightens specialist treatment by inducing specialists to treat only high-severity patients, as opposed to all patients. Tightening specialist treatment is welfare improving if* $-\Delta NB(\underline{s}) > b(\underline{s})\,(1-\omega)$. *In this case, an equity-efficiency trade-off arises only if the GP refers high-severity patients. If the GP refers all patients, the policy increases efficiency but does not affect health inequities.*

The policy of tightening access to specialist treatment involves two possible transitions.[16] Consider a health system where the GP referral system is already tightened, but the specialists have an incentive to treat all patients, which corresponds to scenario 3) in Figure 1. Then, inducing the specialists to treat only high-severity patients, i.e. a tightening of access to specialist services, implies a move from scenario 3) to 1), which again increases allocative efficiency but increases health inequities. An equity-efficiency trade-off arises. Instead, a tightening of access to

[14] By using the expressions of inequalities in specialist treatment from Section 4 and 5, and collecting terms due to income, we get: $\Delta v(\overline{s},\overline{s}) = \overline{\Lambda}_H\left(\overline{\delta}_H - \overline{\delta}_L\right)$, $\Delta v(\overline{s},all) = \left(\overline{\delta}_H - \overline{\delta}_L\right)\overline{\Lambda}_H - (\underline{\delta}_H - \underline{\delta}_L)\underline{\Lambda}_H$, $\Delta v(all,\overline{s}) = \Delta v(all,all) = 0$.

[15] This follows since $sign(\Delta\mathcal{B}(\overline{s},\overline{s}) - \Delta\mathcal{B}(\overline{s},all)) = sign\left[b(\underline{s})\omega + (B(\underline{s}) - b(\underline{s}))\right] > 0$.

[16] Tightening specialist treatment is welfare improving if $\Delta W(\overline{s},\overline{s}) > \Delta W(\overline{s},all)$ or $\Delta W(all,\overline{s}) > \Delta W(all,all)$. Both these inequalities are satisfied when $-\Delta NB(\underline{s}) > b(\underline{s})\,(1-\omega)$.

specialist services, a move from scenario 4) to 2), will increase allocative efficiency, but does not affect health inequities. For these results to hold the condition $-\Delta NB(\underline{s}) > b(\underline{s})(1-\omega)$ has to be satisfied, as this condition ensures that tightening specialist treatment is welfare improving. The condition holds when the welfare loss for a low-severity patient from being treated by specialist is higher, in absolute value (recall $\Delta NB(\underline{s}) < 0$), than the patient health loss due to the delay in treatment from being sent back to the GP by the specialist. This condition is always satisfied if the health loss due to the delay is sufficiently small.

Proposition 8 *Consider a policy that tightens the referral system by inducing GPs to refer only high-severity patients, as opposed to all patients.*
i) Suppose the specialist treats only high-severity patients. Tightening the referral system is welfare improving if $b(\underline{s})(1-\omega) > \frac{\overline{\lambda}_H(1-\overline{\delta}_H)+\overline{\lambda}_L(1-\overline{\delta}_L)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta NB(\overline{s})$, *and an equity-efficiency trade-off arises.*
ii) Suppose the specialist treats all patients. Tightening of the referral system is welfare improving if $-\Delta NB(\underline{s}) > \frac{\overline{\lambda}_H(1-\overline{\delta}_H)+\overline{\lambda}_L(1-\overline{\delta}_L)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta NB(\overline{s})$, *and again an equity-efficiency trade-off arises.*

The policy of tightening the referral system also involves two possible transitions, depending on whether the specialist treats only high-severity patients or all types of patients. Consider a health system with a weak GP referral system where the GP refers all patients and specialists treat only high-severity patients, which corresponds to scenario 2). Then, inducing the GP to refer only patients with high-severity signal, which corresponds to a tightening of the referral system, implies a move from scenario 2) to scenario 1). This transition is welfare improving if $b(\underline{s})(1-\omega) > \frac{\overline{\lambda}_H(1-\overline{\delta}_H)+\overline{\lambda}_L(1-\overline{\delta}_L)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta NB(\overline{s})$.[17] Given that the specialist sends low-severity patients to the GP, this policy is welfare improving only if the delay for low-severity patients in getting treatment is sufficiently high relative to the frequency of the mistakes that the GP does in treating the high-severity patients. If this condition holds, then tightening the GP referral system increases allocative efficiency but also increases health inequities, generating an equity-efficiency trade-off.

Second, consider a health system with a weak GP referral system, and specialists have incentives to treat all patients, which is described under scenario 4). Then, tightening the GP

[17] This inequality follows from $\Delta W(\overline{s},\overline{s}) > \Delta W(all,\overline{s})$.

referral system, a move from scenario 4) to 3), increases welfare if $-\Delta NB(\underline{s}) > \frac{\overline{\lambda}_H(1-\overline{\delta}_H)+\overline{\lambda}_L(1-\overline{\delta}_L)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta NB(\overline{s})$.[18] This condition requires that the welfare loss of those high-severity patients for which GP observes low severity, which happens infrequently, is lower than the welfare loss for the low-severity patients correctly diagnosed by the GP, which happens frequently, but are treated by the specialist. This condition is satisfied if the GP makes sufficiently few mistakes when diagnosing a high severity ($\overline{\delta}_H, \overline{\delta}_L$ are sufficiently high), and this is further reinforced the larger is the difference in the cost between specialist and GP treatment ($(C(\overline{s}) - c)$ is large). If this condition holds, then tightening the GP referral system increases allocative efficiency but also increases health inequities, generating an equity-efficiency trade-off.

The key insight is that whenever introducing a tighter referral system is welfare improving, then this policy increases allocative efficiency but also increases health inequities, generating an equity-efficiency trade-off.

7 Conclusions

To address the financial sustainability of health spending, policymakers regularly introduce new policies that aim at containing costs without harming quality of care. Two policies that have been used to contain costs relate to the interface between primary and secondary care providers. One policy is to tighten the gatekeeping role of primary care providers (GPs) to induce them to refer only the more severe patients to secondary care providers (hospital specialists). A second policy is to tighten access to specialist services to ensure that this more expensive type of care is only available to more severe patients, with less severe patients being treated instead by primary care providers.

This study has provided a theoretical framework to assess these policy interventions and has investigated whether such policies generate an equity-efficiency trade-off, introducing a tension between the ubiquitous policy objective of reducing health inequalities and improving the allocative efficiency of health systems. In our model a purchaser has contracts with two providers of health services, a GP and a hospital specialist, who are reimbursed based on common payment systems: the GP is paid either by fee-for-service, capitation or a combination of the two, and the

[18] This inequality follows from $\Delta W(\overline{s}, all) > \Delta W(all, all)$.

hospital specialist is financed through a DRG-based payment system. Patients differ in severity and in socioeconomic status, and the GP receives an informative signal on the severity of the patient following an examination, which is more informative for patients with higher socioeconomic status, for example because these patients are better able to describe their symptoms.

We generally find that health inequities are higher in health systems with tighter referrals where the GP refers only high-severity patients. Instead, health inequalities are smaller in health systems where specialists have stronger incentives to treat patients. In relation to policies, we show that a tightening of a GP referral system, generally increases allocative efficiency but also increases health inequities. A tightening of access to specialist services increases allocative efficiency and health inequities when the GP refers only severe patients, but has no effect on health inequities when the GP refers all patients. These results suggest that an equity-efficiency trade-off is likely to arise in several circumstances.

Given the current economic climate following the COVID-19 pandemic, cost containment policies are likely to become more prevalent. There may therefore be scope for investigating whether such equity-efficiency trade-offs arise within other contexts in the health sector. There may also be scope for additional empirical evidence. The empirical literature has well documented the presence of inequalities in health and healthcare utilisation. But there is little work that looks at the equity implications of introducing cost containment policies, and in particular policies that aim at reducing referrals and containing access to specialists, possibly because these policies are introduced at a national level making causal identification difficult. Our analysis provides some testable hypotheses that could be the subject of future empirical work.

References

[1] Allard, M., Jelovac, I., Léger, P.T., 2011. Treatment and referral decisions under different physician payment mechanisms. Journal of Health Economics 30, 880–893.

[2] Bago d'Uva, T., Jones, A.M., 2009. Health care utilisation in Europe: New evidence from the ECHP. Journal of Health Economics 28, 265–279.

[3] Bago d'Uva, T., Jones, A.M., van Doorslaer, E., 2009. Measurement of horizontal inequity in health care utilisation using European panel data. Journal of Health Economics 28, 280–289.

[4] Brekke, K.R., Holmås, T.H., Monstad, K., Straume, O.R., 2018. Socio-economic status and physicians' treatment decisions. Health Economics 27, e77–e89.

[5] Brekke, K.R., Nuscheler, R., Straume, O.R., 2007. Gatekeeping in health care. Journal of Health Economics 26, 149–170.

[6] Brekke, K.R., Siciliani, L. and Straume, O.R., 2011. Hospital competition and quality with regulated prices. Scandinavian Journal of Economics 113(2), 444-469.

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

[8] Chen, A., Lakdawalla, D., 2019. Healing the Poor: The Influence of Patient Socioeconomic Status on Physician Supply Responses. Journal of Health Economics 64, 43–54.

[9] Devaux, M., 2015. Income-related inequalities and inequities in health care services utilisation in 18 selected OECD countries. The European Journal of Health Economics 16, 21–33.

[10] Deveugele, M., Derese, A., De Maesschalck, S., Willems, S., Van Driel, M., De Maeseneer, J., 2005. Teaching communication skills to medical students, a challenge in the curriculum? Patient Education and Counseling 58, 265–270.

[11] Ellis, R.P. and McGuire, T.G., 1986. Provider behavior under prospective reimbursement: Cost sharing and supply. Journal of Health Economics 5(2), 129-151.

[12] Glazer, A., 2004. Motivating devoted workers. International Journal of Industrial Organization 22, 427–440.

[13] Godager, G., Iversen, T., Ma, C. to A., 2015. Competition, gatekeeping, and health care access. Journal of Health Economics 39, 159–170.

[14] González, P., 2010. Gatekeeping versus direct-access when patient information matters. Health economics 19(6), 730-754.

[15] Greenfield, S., Kaplan, S.H., Ware, J.E., Yano, E.M., Frank, H.J.L., 1988. Patients' participation in medical care. Journal of General Internal Medicine 3, 448–457.

[16] Griebenow, M., Kifmann, M., 2021. Diagnostics and Treatment: On the Division of Labor between Primary Care Physicians and Specialists, HcHe Research Paper.

[17] Kaarboe, O., Siciliani, L., 2011. Multi-tasking, quality and pay for performance. Health Economics 20, 225–238.

[18] Malcomson, J.M., 2004. Health Service Gatekeepers. The RAND Journal of Economics 35, 401–421.

[19] Mariñoso, B.G., Jelovac, I., 2003. GPs' payment contracts and their referral practice. Journal of Health Economics 22, 617–635.

[20] van Doorslaer, E., Koolman, X., Jones, A.M., 2004. Explaining income-related inequalities in doctor utilisation in Europe. Health Economics 13, 629–647.

[21] van Doorslaer, E., Masseria, C., 2004. Income-Related Inequality in the Use of Medical Care in 21 OECD Countries. OECD Health Working Papers.

8 Appendix

8.1 Appendix A1. Proof of Proposition 1.

$\overline{p} > \underline{p}$ if

$$\begin{aligned} & \frac{\alpha^{gp}\left(B(\overline{s})-b(\overline{s})\right)\Pr(\,s=\overline{s}|\,\sigma=\overline{s})-\left(\alpha^{gp}b(\underline{s})\left(1-\omega\right)+k\right)\Pr(\,s=\underline{s}|\,\sigma=\overline{s})}{\left(1-\Pr(\,s=\underline{s}|\,\sigma=\overline{s})\right)} \\ > \quad & \frac{\alpha^{gp}\left(B(\overline{s})-b(\overline{s})\right)\Pr(\,s=\overline{s}|\,\sigma=\underline{s})-\left(\alpha^{gp}b(\underline{s})\left(1-\omega\right)+k\right)\Pr(\,s=\underline{s}|\,\sigma=\underline{s})}{\left(1-\Pr(\,s=\underline{s}|\,\sigma=\underline{s})\right)}. \end{aligned} \tag{38}$$

Collecting terms, we obtain:

$$\begin{aligned} & \alpha^{gp}\left(B(\overline{s})-b(\overline{s})\right)\left[\frac{\Pr(\,s=\overline{s}|\,\sigma=\overline{s})}{\left(1-\Pr(\,s=\underline{s}|\,\sigma=\overline{s})\right)}-\frac{\Pr(\,s=\overline{s}|\,\sigma=\underline{s})}{\left(1-\Pr(\,s=\underline{s}|\,\sigma=\underline{s})\right)}\right] \\ +\left(\alpha^{gp}b(\underline{s})\left(1-\omega\right)+k\right)\left[\frac{\Pr(\,s=\underline{s}|\,\sigma=\underline{s})}{\left(1-\Pr(\,s=\underline{s}|\,\sigma=\underline{s})\right)}-\frac{\Pr(\,s=\underline{s}|\,\sigma=\overline{s})}{\left(1-\Pr(\,s=\underline{s}|\,\sigma=\overline{s})\right)}\right] \quad > \quad 0 \end{aligned} \tag{39}$$

Notice that, for a given signal σ, the severity is either low or high, e.g. $(1-\Pr(\,s=\underline{s}|\,\sigma=\overline{s}))=\Pr(\,s=\overline{s}|\,\sigma=\overline{s})$. Hence, the first line of (39) is zero. Substituting the relevant probabilities in the second line of (39), after some rearrangements, we obtain

$$\begin{aligned} & \left(\alpha^{gp}b(\underline{s})\left(1-\omega\right)+k\right) \\ \times\left(\frac{\underline{\lambda}_L\underline{\delta}_L+\underline{\lambda}_H\underline{\delta}_H}{\underline{\lambda}_L(1-\underline{\delta}_L)+\underline{\lambda}_H(1-\underline{\delta}_H)}-\frac{\overline{\lambda}_L(1-\overline{\delta}_L)+\overline{\lambda}_H(1-\overline{\delta}_H)}{\overline{\lambda}_L\overline{\delta}_L+\overline{\lambda}_H\overline{\delta}_H}\right) \quad > \quad 0, \end{aligned}$$

where the term in the first line is positive by assumption, and the first (second) term of the second line is larger (smaller) than one. ■

8.2 Appendix A2. Income-related health inequalities. Equation (21).

The expected benefit from treatment for each income group is:

$$\mathcal{B}_i=\overline{\Lambda}_i\left[\overline{\delta}_iB(\overline{s})+(1-\overline{\delta}_i)b(\overline{s})\right]+\underline{\Lambda}_ib(\underline{s})\left[1-(1-\underline{\delta}_i)\left(1-\omega\right)\right],i=L,H \tag{40}$$

By adding and subtracting the terms in line 6, we obtain:

$$\begin{aligned}\mathcal{B}_H - \mathcal{B}_L &= \overline{\Lambda}_H \left[\overline{\delta}_H B(\overline{s}) + (1-\overline{\delta}_H) b(\overline{s})\right] + \underline{\Lambda}_H b(\underline{s}) \left[1 - (1-\underline{\delta}_H)(1-\omega)\right] \\ &\quad -\overline{\Lambda}_L \left[\overline{\delta}_L B(\overline{s}) + (1-\overline{\delta}_L) b(\overline{s})\right] - \underline{\Lambda}_L b(\underline{s}) \left[1 - (1-\underline{\delta}_L)(1-\omega)\right]\end{aligned} \tag{41}$$

By adding and subtracting the terms in second and fourth line we obtain

$$\begin{aligned}\mathcal{B}_H - \mathcal{B}_L &= \overline{\Lambda}_H \left[\overline{\delta}_H B(\overline{s}) + (1-\overline{\delta}_H) b(\overline{s})\right] + \underline{\Lambda}_H b(\underline{s}) \left[1 - (1-\underline{\delta}_H)(1-\omega)\right] \\ &\quad +\overline{\Lambda}_H \overline{\delta}_L \left(B(\overline{s}) - b(\overline{s})\right) - \overline{\Lambda}_H \overline{\delta}_L \left(B(\overline{s}) - b(\overline{s})\right) \\ &\quad -\overline{\Lambda}_L \left[\overline{\delta}_L B(\overline{s}) + (1-\overline{\delta}_L) b(\overline{s})\right] - \underline{\Lambda}_L b(\underline{s}) \left[1 - (1-\underline{\delta}_L)(1-\omega)\right] \\ &\quad +\underline{\Lambda}_H \underline{\delta}_L b(\underline{s})(1-\omega) - \underline{\Lambda}_H \underline{\delta}_L b(\underline{s})(1-\omega) \\ &= \overline{\Lambda}_H \left(\overline{\delta}_H - \overline{\delta}_L\right) \left[B(\overline{s}) - b(\overline{s})\right] + \underline{\Lambda}_H \left(\underline{\delta}_H - \underline{\delta}_L\right) b(\underline{s})(1-\omega) \\ &\quad - \left(\overline{\Lambda}_L - \overline{\Lambda}_H\right) \left[\overline{\delta}_L B(\overline{s}) + (1-\overline{\delta}_L) b(\overline{s})\right] - \left(\underline{\Lambda}_L - \underline{\Lambda}_H\right) b(\underline{s}) \left[\underline{\delta}_L - (1-\underline{\delta}_L)\omega\right].\end{aligned} \tag{42}$$

■

8.3 Appendix A3. Proof of Proposition 4.

$\widetilde{p} > \underset{\sim}{p}$ if

$$\begin{aligned}&\alpha^{gp} \left(B(\overline{s}) - b(\overline{s})\right) \left[\Pr(s = \overline{s} | \sigma = \overline{s}) - \Pr(s = \overline{s} | \sigma = \underline{s})\right] \\ &-\alpha^{gp} \left(B(\underline{s}) - b(\underline{s})\right) \left[\Pr(s = \underline{s} | \sigma = \underline{s}) - \Pr(s = \underline{s} | \sigma = \overline{s})\right] > 0.\end{aligned} \tag{43}$$

This is the case given our assumption that $(B(\overline{s}) - b(\overline{s})) > (B(\underline{s}) - b(\underline{s})) > 0$, and since

$$\Pr(s = \overline{s} | \sigma = \overline{s}) - \Pr(s = \overline{s} | \sigma = \underline{s}) = \Pr(s = \underline{s} | \sigma = \underline{s}) - \Pr(s = \underline{s} | \sigma = \overline{s}). \tag{44}$$

To see the last statement, notice that the sum of the probabilities over types for a given signal must equal one (i.e. for a given signal the patient is either low or high severity). This holds true for both signals. Hence the result is obtained. ■

Figure 1. GP and specialist referral and treatment scenarios

		Specialist	
		Treats high-severity and *refers* back low-severity patients	*Treats* high-severity and low-severity patients
GP	*Refers* high-severity and *treats* low-severity patients	Scenario 1. $\underline{p} < p < \bar{p}$ Pro-rich inequities in specialist treatment and health benefit	Scenario 3. $\underset{\sim}{p} < p < \tilde{\bar{p}}$ Pro-rich (pro-poor) inequities in specialist treatment and health benefit if the low-severity incidence is sufficiently small (large)
	Refers high-severity and low-severity patients	Scenario 2. $p < \underline{p}$ No inequities in treatments and health benefits	Scenario 4. $p < \underset{\sim}{p}$ No inequities in treatments and health benefits

Note: p is the fee received by the GP for each patient visit.