Journal of Health Economics 87 (2023) 102715

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/jhe

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

Oddvar Kaarboe [a,b,c,*], Luigi Siciliani [d]

[a] IGS, University of Bergen, Norway
[b] Department of Economics, University of Bergen, Norway
[c] HELED, University of Oslo, Norway
[d] Department of Economics and Related Studies, University of York, Heslington, York, UK

ARTICLE INFO

JEL classification:
I11
I14
I18

Keywords:
Primary care
Secondary care
Equity
Payment system
Allocative efficiency
Inequalities
Access

ABSTRACT

We analyse how payment systems for general practitioners (GPs) and hospital specialists affect inequalities in healthcare treatments, referrals, and patient health. We present a model of contracting with two providers, a GP and a hospital specialist, with patients differing in severity and socioeconomic status, and the GP only receiving an informative signal on severity. We investigate four health system configurations depending on whether the GP refers and the specialist treats only high-severity patients or patients with any severity. We show that an increase in the GP fee, which induces GPs to refer only high-severity patients, increases utilitarian welfare but also increases inequities in access to specialist visits. A reduction in the DRG reimbursement to hospital specialists, which induces specialists to treat only high-severity patients, increases utilitarian welfare but also increases inequities in access to specialist visits when the GP refers only high-severity patients.

1. Introduction

Reductions in health and healthcare inequalities are ubiquitous policy objectives. Despite these objectives, inequalities in healthcare utilisation persist. For specialist visits, the empirical evidence suggests that, for a given need, individuals with higher socioeconomic status have better access to specialist visits in most OECD countries (van Doorslaer et al., 2004; van Doorslaer and Masseria, 2004; Bago d'Uva and Jones, 2009; Devaux, 2015). For general practitioner (GP) visits, the results are mixed, with some evidence suggesting that individuals with lower socioeconomic status have more GP visits, for a given level of need, in a sub-set of countries (van Doorslaer and Masseria, 2004; Bago d'Uva et al., 2009).

In this study, we address two research questions. First, how do different payment systems for GPs and hospital specialists affect inequalities in GP and specialist visits? Second, do policies aimed at increasing utilitarian welfare (defined as patient benefits net of provider costs) increase inequities in specialist visits, and therefore in patient health? 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 for GPs and specialists generate different degrees of inequities in treatments and referrals, and health inequities. We consider two policies aimed at increasing utilitarian welfare or more broadly the (allocative) efficiency of health systems: a policy that incentivises GPs

☆ We thank two anonymous referees and the editor for very helpful comments and suggestions. The paper is partly funded by the Research Council of Norway (288592).

* Corresponding author at: Department of Economics, University of Bergen, Norway.
E-mail addresses: oddvar.kaarboe@uib.no (O. Kaarboe), luigi.siciliani@york.ac.uk (L. Siciliani).

https://doi.org/10.1016/j.jhealeco.2022.102715
Received 7 December 2021; Received in revised form 23 November 2022; Accepted 26 November 2022
Available online 7 December 2022

to refer only high-severity patients to specialists and treat low-severity patients, and a policy that incentivises specialists to treat only high-severity patients.

To answer our research questions 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 their symptoms to the GP.

Based on the signal, the GP decides either to treat the patient, or to refer to a specialist. The specialist observes the severity of each referred patient and decides whether to treat or refer the patient back to the GP. These assumptions give rise to four possible health system configurations: (1) the GP refers only patients with high-severity signal and the specialist treats only high-severity patients; (2) the GP refers all patients, but the specialist treats only high-severity patients; (3) the GP refers only patients with high-severity signal, and the specialist treats all patients; (4) 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 reimbursement system. Both the GP and the hospital specialist are altruistic and obtain utility both from patients' benefits of treatments and income.[1] We assume that the GP treatment cost is independent of severity (e.g. drug treatment), but that the specialist treatment cost increases with 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 utility is reduced.

Our key findings are as follows. We generally find that inequities in access to specialist visits are higher in health systems where the GP refers only high-severity patients, and lower in systems where GPs refer both high- and low-severity patients. More precisely, inequities in access to specialist visits are highest under scenario (1) when the GP refers patients with high-severity signal and the specialist treats only high-severity patients. Inequities are intermediate in scenario (3) when the GP refers patients with high-severity signal and the specialist treats all patients. When the GP refers all patients under scenarios (2) and (4), there are no inequities in access regardless of whether the specialist treats only high-severity patients or all patients. The intuition is as follows. Whenever the GP refers only patients with high-severity signal, patients with low socioeconomic status are less likely to be referred to a specialist when they have high severity because their ability to convey the high-severity signal is lower. Patients with low socioeconomic status are instead more likely to be referred to a specialist if they have low severity because the low-severity signal is also less informative. If the specialist treats only high-severity patients, then all patients with low-severity signal are sent back to the GP by the specialist, regardless of their socioeconomic status, which generates inequities in access and health inequities in favour of patients with high socioeconomic status. Instead, if the specialist treats all patients, then low severity patients with low socioeconomic status are more likely to benefit from specialist treatment, which reduces health inequities in favour of the rich (or, if the incidence of low-severity patients is high, can generate a pro-poor health gradient). Whenever the GP refers all patients, all patients with low severity are either sent back to the GP, if the specialist only treats high-severity patients, or treated by the specialist, regardless of their socioeconomic status, which eliminates health inequities.

We show that utilitarian welfare (allocative efficiency), which is given by patient benefits net of providers' costs, is highest when the GP refers more selectively and the specialist only treats high-severity patients. Utilitarian welfare is instead lowest when the GP refers all patients and the specialist has incentives to treat all patients.

We characterise the effect of policies that induce GPs to refer more selectively only the high-severity patients, or induce specialists to restrict the access to specialist services. These policies are regularly discussed as interventions to contain costs and improve the efficiency and sustainability of health spending. For example, GPs could be incentivised to treat low-severity patients rather than referring them to a specialist by increasing the FFS fee paid to GPs. Instead, specialists could have a weaker incentive to treat patients if the DRG reimbursement is sufficiently low.

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). An increase in the GP fee induces the GP to refer only patients with a high-severity signal and implies a move from scenario (2) to scenario (1). The introduction of a more selective referral system increases utilitarian welfare but also increases inequities in access to specialist services and health, therefore generating a trade-off between equity in access and allocative efficiency.

Similarly, consider a health system where the GP refers patients more selectively, but specialists have an incentive to treat all patients, which corresponds to scenario (3). Restricting access to specialist services, for example through a reduction in the DRG reimbursement to the hospital specialist, implies a move from scenario (3) to (1). This increases utilitarian welfare, but also increases inequities in access to specialist services and health inequities.

Finally, consider a health system with a loose GP referral system, where the GP refers all patients, and specialists have incentives to treat all patients, which is described under scenario (4). Then, requiring GPs to refer more selectively, a move from scenario (4) to (3), increases utilitarian welfare but also increases inequities in access to specialist services and health inequities. Again, a trade-off between equity in access and allocative efficiency arises. Instead, restricting access to specialist services, a move from scenario (4) to (2), will increase utilitarian welfare, but has no effect on inequities in access to specialist services.[2]

[1] 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).

[2] It is unlikely that policymakers would move from scenario (2) to (3), or from (3) to (2). The former would require GPs to refer more selectively and at the same time ease access to specialist services. The latter would require GPs to refer less selectively. We therefore do not discuss these scenarios.

In summary, a trade-off between allocative efficiency, as measured by utilitarian welfare, and equity in access to specialist services is likely to arise in several circumstances. Our analysis is positive rather than normative. Rather than deriving an optimal payment system that maximises a welfare function, we instead investigate the effects of realistic policy interventions, emphasising possible trade-offs that may arise as a result.

An important assumption behind our results on inequalities in access is that patients with higher socioeconomic status can communicate more easily, which improves the accuracy of the signalling of severity. This is a realistic assumption as research shows that the physician–patient communication tends to differ according to the socioeconomic background of the patient, see e.g. the systematic review by Deveugele et al. (2005) and the meta analysis by Verlinde et al. (2012). The gradient may arise either because physicians provide less information to patients with lower socioeconomic status, because they think that these patients are less interested in learning about their health or are less able to understand this information (Waitzkin, 1985; Street, 1991; Starfield, 2006; Baron-Epel et al., 2007; Cerin and Leslie, 2008; Williams et al., 2010), or because patients communicate differently with their doctor depending on their socioeconomic status (Verlinde et al., 2012). As a result, patients with a lower socioeconomic status are significantly less involved in treatment decisions, are approached in a more directive way during the consultation, and are less frequently asked to take responsibility for care than patients with a higher socioeconomic status (Verlinde et al., 2012).

The effects of good communication have been studied in Greenfield et al. (1988) and in Tavakoly Sany et al. (2020). In the first study, patients with diabetes were randomised to a pre-visit coaching session, at which 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 the non-coached control groups, 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. In the second study, physicians and hypertensive patients were randomly allocated to the intervention and control groups, and physicians in the intervention group received educational training. The control group received the routine care. The primary outcome was a reduction in systolic and diastolic blood pressure from baseline to 6 months. The secondary outcome was promoting health literacy skills in hypertensive patients. The authors find a significant improvement in physician–patient communication skills, hypertension outcomes, medication adherence, and self-efficacy among the patients being managed by the physicians receiving training, compared to the control group.

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 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 utilitarian welfare analysis. 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 healthcare providers, when doctors cannot observe severity directly, but only 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 speciality care, and that fundholding induces almost as many 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 the 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 diagnostic 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 GP 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' incentives to refer patients to a specialist, and show that the effect is in principle indeterminate. On one hand, competition induces the GP to refer more often in order to improve patient satisfaction. On the other hand, they tend to earn more by treating patients themselves, thus 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 specialisation in the secondary care market. They show that hospital competition is amplified by higher GP attendance but dampened by improved diagnostic 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. They show that patients in Norway with diabetes (type II) with low education have shorter consultations but more medical tests. Instead, patients with low income have shorter consultations and fewer medical tests. 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, and therefore their income and socioeconomic status. Using Medicare reimbursement changes, they find that physicians facing an increase in reimbursement rates increase utilisation more for richer, relative to poorer, patients.[3] We differ from these

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

studies by using an informative signal framework, by allowing a more explicit interaction between the GP and the specialist, and by investigating the implications of different policy interventions and possible tensions between equity in access and allocative efficiency as captured by utilitarian welfare.

3. The model

We present a model of provider behaviour with a GP and a hospital specialist serving a population of patients, which is normalised to one. Patients have high or low severity, $s \in \{\underline{s}, \overline{s}\}$, and high and low income[4], $i \in \{L, H\}$, 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, Hungary, Netherlands, New Zealand, Poland, Portugal, Spain and United Kingdom. Many Health Maintenance Organisations in the US also have gatekeeping physicians (Glied, 2000).[5] 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 the severity of their condition. 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. 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 $\Pr_i(\sigma = \overline{s}|\, s = \overline{s}) = \overline{\delta}_i > 0.5$, and similarly for a *low* -severity signal conditional on the patient being *low* severity, $\Pr_i(\sigma = \underline{s}|\, s = \underline{s}) = \underline{\delta}_i > 0.5$. Therefore the signal is informative.[6]

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 (see online appendix A1), 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 $\underline{\delta}_H > \underline{\delta}_L$. This implies that, for given severity, the signal is more informative for patients with high income because patients and doctors communicate better, and this facilitates the assessment of the health state of the patient. Moreover, we assume that $\overline{\delta}_i > \underline{\delta}_i$: 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 coughing for more than six weeks (lung cancer), lumps in the breast (breast cancer), pain level, fever, and unintended weight loss, are more likely to be detected by the doctor. However, the absence of such symptoms does not provide a strong signal of low severity. For example, lung cancer often has no symptoms until it has spread (metastasised) since there are few specialised nerves (pain receptors) in the lungs (Harle et al., 2014), and some types of breast cancer (e.g. invasive lobular carcinoma and inflammatory breast cancer) are less likely to cause breast lumps.

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, $B(\underline{s}) \geq b(\underline{s})$. Specialists spend several years training in a specific field of medicine, such as orthopaedics or ophthalmology. This higher level of training allows them not only to determine the level of severity of the patient but also to make a more accurate diagnosis of the underlying health problem and to recommend a treatment that is tailored towards patients' need, therefore increasing the expected health benefit. Our assumption is that the specialist can do

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

[5] The review by Jelovac (2014) summarises the empirical evidence on the effects of gatekeeping. The evidence is mixed. It suggests that gatekeeping decreases patients' satisfaction, but is significantly associated with a lower utilisation of health services and lower expenditure.

[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 $\Pr_i(\sigma = \overline{s}|\, s = \underline{s}) = (1 - \underline{\delta}_i)$.

at least as well as the GP in improving patients' health. For some health conditions, low-severity patients can still be treated with a standardised treatment, such as a drug, in which case the patient's benefit from being treated by a GP or a specialist is the same.[7] Last, we assume that high-severity patients benefit more from being treated by a specialist, $B(\overline{s}) - b(\overline{s}) > B(\underline{s}) - b(\underline{s})$.

Providers' cost. We assume that GP treatment cost is c, and is independent of severity (e.g. drug treatment). Specialist treatment cost is equal to $C(s)$, increases with severity, and is more expensive than GP treatment, $C(\overline{s}) > C(\underline{s}) > c$.[8]

Specialist utility function. We assume that specialists can always diagnose patient severity with no mistakes.[9] After diagnosis, the hospital specialist has two choices, either to treat the patients or to refer them back to the GP. If the specialist treats a patient with severity s, her utility, defined with $V(\cdot)$, is given by

$$V(treat, s) = \begin{cases} T + P(s) - C(s) + \alpha^h B(s) & if \quad s = \overline{s} \\ T + P(s) - C(s) + \alpha^h B(s) - \Omega & if \quad s = \underline{s} \end{cases} \tag{1}$$

where $P(s)$ is a DRG-reimbursement 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 and 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 about treating a patient who 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). This 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. The parameter Ω plays an important role in the model. As shown in Section 4, although specialists have a stronger altruistic benefit from treating high-severity patients, the utility from treating a low-severity patient is always positive as long as the price mark-up is positive. However, hospitals in several health systems have stringent capacity constraints, and these limit the ability of specialists to treat low-severity patients. Introducing the parameter Ω is a simple way to capture such capacity constraints and distinguish between health systems. 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:

$$V(referback, s) = T + \alpha^h \omega b(s), \quad \text{with } s \in \left\{\underline{s}, \overline{s}\right\}, \tag{2}$$

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.[10]

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

$$\begin{aligned} \Delta V(\overline{s}) &= V(treat, \overline{s}) - V(referback, \overline{s}) = P(\overline{s}) - C(\overline{s}) + \alpha^h \left(B(\overline{s}) - \omega b(\overline{s})\right), \\ \Delta V(\underline{s}) &= V(treat, \underline{s}) - V(referback, \underline{s}) = P(\underline{s}) - C(\underline{s}) + \alpha^h \left(B(\underline{s}) - \omega b(\underline{s})\right) - \Omega \end{aligned} \tag{3}$$

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 out the two latter (unlikely) scenarios by making the following assumptions.

Assumption **A1**: $\Delta V(\overline{s}) > 0$. This assumption ensures that the specialist always has an incentive to treat a high-severity patient rather than referring the patient back to the GP, or more extensively, $\alpha^h B(\overline{s}) + P(\overline{s}) > C(\overline{s}) + \alpha^h \omega b(\overline{s})$. The sum of the non-monetary patient benefits and the monetary ones, given by the DRG reimbursement tariff, is larger than the treatment cost and the non-monetary cost for the patient from delayed GP treatment.

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

$$\alpha^h \left(B(\overline{s}) - B(\underline{s})\right) + \Omega + P(\overline{s}) - P(\underline{s}) > C(\overline{s}) - C(\underline{s}) + \alpha^h \omega \left(b(\overline{s}) - b(\underline{s})\right). \tag{4}$$

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 reimbursed DRG reimbursement 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 $s \in \left\{\underline{s}, \overline{s}\right\}$ is given by

$$U(treat, s) = t + p - c + \alpha^{gp} b(s), \tag{5}$$

[7] Note that if the patient had lower health benefit when treated by a specialist due, for example, to over-treatment, our results would be qualitatively the same, but the parameter space over which the specialist treats only high-severity patients is expanded (see Eq. (7) in Section 4).

[8] Assuming that the GP treatment cost also increases with severity would not qualitatively change our results. See online appendix A6 for details.

[9] 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.

[10] We assume that the delay does not affect patient severity. Allowing for the possibility that the delay increases severity would add complexity to the model, but would not alter the main insights. If a less severe patient becomes more severe due to the delay (for example with some probability), the specialist will never send her back to the GP. This will shift such cases from scenario 1 and 2 (where the specialist refers back the patient) to scenario 3 and 4 (the specialist treats all cases), but conceptually the analyses are unchanged.

		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

Fig. 1. GP and specialist referral and treatment scenarios. Note: p is the fee received by the GP for each patient visit.

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

$$U(refer, s) = \begin{cases} t + \alpha^{gp} B(s) & if \quad s = \overline{s} \\ \alpha^{gp} \omega b(s) + t + p - c - k & if \quad s = \underline{s} \end{cases} \tag{6}$$

where $k \gtreqless 0$ captures a potential financial penalty for (inappropriately) referring a low-severity patient to the specialist.[11]

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 respectively in Sections 3 and 4. For each scenario relating to the specialist behaviour, we distinguish two further sub-cases regarding the GP behaviour, whether the GP refers only high-severity patients to the specialist, or both high- and low-severity patients. This produces four scenarios (Fig. 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 high- and low-severity (scenario 4).

[11] 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, $k = -p$.

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:

$$\Delta V(\underline{s}) = P(\underline{s}) - C(\underline{s}) + \alpha^h \left(B(\underline{s}) - \omega b(\underline{s}) \right) < \Omega. \tag{7}$$

This condition holds when the DRG reimbursement tariff for a 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 reimbursement tariff covers only a proportion of the costs (e.g. 30%–60%). In other systems, there may be penalties when hospitals admit a 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 reimbursement 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. In turn, this 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, taking into account that low-severity patients will be sent back to the GP if referred to the specialist. The GP maximises 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:

$$\mathbf{E}U(refer, \sigma) = t + \alpha^{gp} B(\overline{s}) \Pr(s = \overline{s} | \sigma) + \left(\alpha^{gp} \omega b(\underline{s}) + p - c - k \right) \Pr(s = \underline{s} | \sigma). \tag{8}$$

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

$$\mathbf{E}U(treat, \sigma) = t + p - c + \alpha^{gp} b(\overline{s}) \Pr(s = \overline{s} | \sigma) + \alpha^{gp} b(\underline{s}) \Pr(s = \underline{s} | \sigma). \tag{9}$$

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

$$\Delta \mathbf{E}U(\sigma) = \alpha^{gp} \left(B(\overline{s}) - b(\overline{s}) \right) \Pr(s = \overline{s} | \sigma) - (\alpha^{gp} b(\underline{s})(1 - \omega) + k) \Pr(s = \underline{s} | \sigma) - (p - c)(1 - \Pr(s = \underline{s} | \sigma))$$

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 the 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

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

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

Suppose that $\alpha^{gp} b(\underline{s})(1 - \omega) + k > 0$, so that $\overline{p} > \underline{p}$ (see online appendix A2). If the fee received by the GP for each patient visit is low, i.e. $p < \underline{p}$, the GP always refers the patient to the specialist. If the fee is intermediate, i.e. $\underline{p} < p < \overline{p}$, 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. 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 Fig. 1, and the case with a low GP fee for a visit corresponds to scenario 2 in Fig. 1.[12] We discuss these two scenarios in turn in the next two 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:[13]

$$R = \Pr(\sigma = \overline{s}) = \overline{\lambda}_L \overline{\delta}_L + \overline{\lambda}_H \overline{\delta}_H + \underline{\lambda}_L (1 - \underline{\delta}_L) + \underline{\lambda}_H (1 - \underline{\delta}_H). \tag{10}$$

[12] Below, we do not discuss the scenario in which 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.

[13] 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 $\Pr(s = \underline{s}) = \underline{\lambda}_L + \underline{\lambda}_H$.

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 $i = L, H$.[14] 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 referral rates are:

$$r_H - r_L = \left(\overline{\delta}_H - \overline{\delta}_L\right)\overline{\Lambda}_H - (\underline{\delta}_H - \underline{\delta}_L)\underline{\Lambda}_H + \overline{\delta}_L\left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) - (1-\underline{\delta}_L)\left(\underline{\Lambda}_L - \underline{\Lambda}_H\right). \tag{11}$$

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).[15]

The income-related inequalities in the proportion of specialist treatment is: $v_H - v_L = \overline{\Lambda}_H\overline{\delta}_H - \overline{\Lambda}_L\overline{\delta}_L$. 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:

$$v_H - v_L = \overline{\Lambda}_H\left(\overline{\delta}_H - \overline{\delta}_L\right) + \left(\overline{\Lambda}_H - \overline{\Lambda}_L\right)\overline{\delta}_L. \tag{12}$$

The first component is due to the lower accuracy in the severity signal that arises in the interaction between the patient and the GP. In turn, this generates differential access to specialist treatment. For short, we refer to this first component as *inequities* in specialist treatment. Instead, we refer to the second term to inequalities as these reflect severity incidence. Within many publicly-funded health systems, healthcare is supposed to be allocated according to 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 when the patient interacts with the GP. We therefore conclude that there are pro-rich inequities in specialist treatment and pro-poor inequities in GP treatment.[16]

The expected benefit from treatment for each income group is

$$\mathcal{B}_i = \overline{\Lambda}_i\left[\overline{\delta}_i B(\overline{s}) + (1-\overline{\delta}_i)b(\overline{s})\right] + \underline{\Lambda}_i b(\underline{s})\left[\underline{\delta}_i + \left(1-\underline{\delta}_i\right)\omega\right],\ i = L, H. \tag{13}$$

This 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 benefits are given by (see online appendix A3):

$$\begin{aligned}\mathcal{B}_H - \mathcal{B}_L = {} & \overline{\Lambda}_H\left(\overline{\delta}_H - \overline{\delta}_L\right)\left[B(\overline{s}) - b(\overline{s})\right] + \underline{\Lambda}_H\left(\underline{\delta}_H - \underline{\delta}_L\right)b(\underline{s})(1-\omega) \\ & - \left(\overline{\Lambda}_L - \overline{\Lambda}_H\right)\left[\overline{\delta}_L B(\overline{s}) + (1-\overline{\delta}_L)b(\overline{s})\right] - \left(\underline{\Lambda}_L - \underline{\Lambda}_H\right)b(\underline{s})\left[\underline{\delta}_L - (1-\underline{\delta}_L)\omega\right].\end{aligned} \tag{14}$$

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

[14] The incidence (and prevalence) varies across diseases. According to the 2020 National Survey on Drug Use and Health, 8.4% of the U.S. adults had at least one major depressive episode in 2020, National Institute of Mental Health (2022). Abdalla et al. (2020) analyse cardiovascular disease prevalence by income level in the US. Using nationally representative data from nine cycles of the National Health and Nutrition Examination Survey (NHANES) between 1999 and 2016, the authors calculate age-standardised prevalence using the 2010 estimates. The overall prevalence of congestive heart failure was 3.4%, angina was 3.0%, heart attack was 4.4%, and stroke was 3.9%.

[15] Abdalla et al. (2020) found that the US top 20% earners had a lower cardiovascular disease (CVD) prevalence than the remainder of the population. The gaps in age-standardised prevalence between the two groups for the higher-severity CVDs conditions (heart attack and stroke) were respectively 2.1% vs 3.7% and 1.3% vs 3.2%. Similarly, the gap for the lower-severity CVD conditions (congestive heart failure and angina) were 0.9% vs 2.8%, and 1.5% vs 2.7%. Using a recent nationwide health survey in Australia, Hashmi et al. (2021) reported mental disorder prevalence rates across different socioeconomic groups. By comparing prevalence rates among individuals in the highest and lowest income quintile the authors found that the gap in the prevalence rates for the depression (a more severe condition) were respectively 12,3% vs. 5,09%. Similar, the gaps for generalised anxiety disorder (a less severe condition) were 9,82% vs. 5,73%. One indication of significant prevalence difference between low- and high socioeconomic individuals is differences in life expectancy, and it is well known that income predicts mortality (Mackenbach et al., 2008). Chetty et al. (2016), using US income tax data from 1999 to 2015, found that higher income is associated with greater longevity throughout the income distribution, and that the gap in life expectancy between the richest 1% and poorest 1% of individuals was almost 15 years for men and about 10 years for women.

[16] 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.

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 this with the following proposition (scenario 1 in Fig. 1).

Proposition 1. *Let the GP fee for a visit be such that $\underline{p} < p < \overline{p}$ so that the GP only refers 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 $v_H - v_L = \overline{\Lambda}_H - \overline{\Lambda}_L$. 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 $g_H - g_L = \underline{\Lambda}_L - \underline{\Lambda}_H$.

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

$$\mathcal{B}_H - \mathcal{B}_L = \left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) B(\overline{s}) + \left(\underline{\Lambda}_H - \underline{\Lambda}_L\right) b(\underline{s})\omega, \tag{15}$$

with the difference in benefit being amplified by the delay ω. We summarise in the following proposition (scenario 2 in Fig. 1).

Proposition 2. *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, and takes into account that low-severity patients will be treated by the specialists if referred (and differently from Section 4, where low-severity patients are sent back to the GP if referred 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:

$$\mathbf{E}U(refer, \sigma) = t + \alpha^{gp}\left[B(\overline{s})\Pr(\overline{s}|\sigma) + B(\underline{s})\Pr(\underline{s}|\sigma)\right]. \tag{16}$$

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

$$\mathbf{E}U(treat, \sigma) = t + p - c + \alpha^{gp} b(\overline{s})\Pr(\overline{s}|\sigma) + \alpha^{gp} b(\underline{s})\Pr(\underline{s}|\sigma). \tag{17}$$

Therefore, the GP refers the patient when

$$\Delta\mathbf{E}U(\sigma) = \alpha^{gp}\left(B(\overline{s}) - b(\overline{s})\right)\Pr(\overline{s}|\sigma) + \alpha^{gp}\left(B(\underline{s}) - b(\underline{s})\right)\Pr(\underline{s}|\sigma) - p + c > 0. \tag{18}$$

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

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

with $\widetilde{p} > \underset{\sim}{p} > 0$ (see online appendix A4). If the GP fee for a visit is low, i.e. $0 \leq p < \underset{\sim}{p}$, the GP always refers the patient to the specialist. If the GP fee is intermediate, i.e. $\underset{\sim}{p} < p < \widetilde{p}$, 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.

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.[17] The case with an intermediate GP fee for a visit corresponds to scenario 3 in Fig. 1, and the case with a low GP fee for a visit corresponds to scenario 4 in Fig. 1. We discuss these two scenarios in turn in the next two Sections 5.1 and 5.2.

[17] Similarly to Section 4, we do not discuss the scenario in which 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.

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. $\underset{\sim}{p} < p < \widetilde{p}$, the total number of referrals R is, as in Section 4.1, $R = \overline{\lambda}_L\overline{\delta}_L + \overline{\lambda}_H\overline{\delta}_H + \underline{\lambda}_L(1-\underline{\delta}_L) + \underline{\lambda}_H(1-\underline{\delta}_H)$, and 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 equal to:

$$r_H - r_L = \left(\overline{\delta}_H - \overline{\delta}_L\right)\overline{\Lambda}_H - (\underline{\delta}_H - \underline{\delta}_L)\underline{\Lambda}_H + \overline{\delta}_L\left(\overline{\Lambda}_H - \overline{\Lambda}_L\right) - (1-\underline{\delta}_L)\left(\underline{\Lambda}_L - \underline{\Lambda}_H\right). \tag{19}$$

The income-related inequalities in GP referrals are identical to the scenario, in which the GP refers only high-severity patients and the specialist refers low-severity patients back to the GP (see Section 4.1, Eq. (11)), 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 $r_H - r_L = v_H - v_L = g_L - g_H$.

We again decompose inequalities in specialist treatment between inequalities due to income (inequities) in the first two terms in (19) and inequalities due to differences in severity incidence in the last two terms in (19). 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

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

and inequalities in health benefit are given by

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

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 this in the following proposition (scenario 3 in Fig. 1), which isolates the gradient due to the accuracy of the signal.

Proposition 3. *Let the GP fee for a visit be intermediate, $\underset{\sim}{p} < p < \widetilde{p}$, so that the GP only refers 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).*

Income-related health inequities are always higher in scenario 1 than in the current scenario 3. This follows by comparing (14) with (21), 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

$$\mathcal{B}_H - \mathcal{B}_L = \left(\overline{\Lambda}_H - \overline{\Lambda}_L\right)B(\overline{s}) + \left(\underline{\Lambda}_H - \underline{\Lambda}_L\right)B(\underline{s}). \tag{22}$$

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

Proposition 4. *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 possible equity-efficiency trade-offs.

6. Utilitarian welfare

We adopt a utilitarian welfare function defined as the difference between patient benefits and provider costs.[18] 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

$$\Delta NB(s) = B(s) - C(s) - (b(s) - c), \tag{23}$$

is positive. In the following, we assume that:

A3 $\Delta NB(\overline{s}) > 0$. The benefit from being treated by a specialist relative to a GP is positive for the high-severity patients.

A4 $\Delta NB(\underline{s}) < 0$. The benefit from being treated by a specialist relative to a GP is negative for the 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 utilitarian welfare under the first-best solution is given by:

$$W^{fb} = \left(\overline{\lambda}_H + \overline{\lambda}_L\right)\left[B(\overline{s}) - C(\overline{s})\right] + \left(\underline{\lambda}_H + \underline{\lambda}_L\right)\left[b(\underline{s}) - c\right]. \tag{24}$$

Using the utilitarian 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 utilitarian 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 utilitarian 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.

After computing $\Delta W(s,s) = W(s,s) - W^{fb}$, straightforward calculations (see online appendix A5) give:

$$\begin{aligned}
\Delta W(\overline{s},\overline{s}) &= -\left(\overline{\lambda}_H\left(1-\overline{\delta}_H\right) + \overline{\lambda}_L\left(1-\overline{\delta}_L\right)\right)\Delta NB(\overline{s}) \\
&\quad - \left(\underline{\lambda}_H(1-\underline{\delta}_H) + \underline{\lambda}_L(1-\underline{\delta}_L)\right) b(\underline{s})(1-\omega), \\
\Delta W(\overline{s},all) &= -\left(\overline{\lambda}_H\left(1-\overline{\delta}_H\right) + \overline{\lambda}_L\left(1-\overline{\delta}_L\right)\right)\Delta NB(\overline{s}) \\
&\quad + \left(\underline{\lambda}_H(1-\underline{\delta}_H) + \underline{\lambda}_L(1-\underline{\delta}_L)\right)\Delta NB(\underline{s}), \\
\Delta W(all,\overline{s}) &= -\left(\underline{\lambda}_H + \underline{\lambda}_L\right) b(\underline{s})(1-\omega), \\
\Delta W(all,all) &= \left(\underline{\lambda}_H + \underline{\lambda}_L\right)\Delta NB(\underline{s}).
\end{aligned} \tag{25}$$

Let us 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 GP's misinterpretation of the signals: A share of the low-severity patients are treated by the specialist,

[18] An alternative and equivalent notion would be to define welfare as the difference between patient benefit net of the payment to the providers, and providers' profit, given by the provider payments net of treatment costs. The welfare from being treated by a specialist is the difference between patient net benefit, $B(s) - P(s) - T$, and specialist profit, $P(s) + T - C(s)$, which gives $B(s) - C(s)$. Similarly, the welfare from being treated by a GP is the difference between patient net benefit, $b(s) - p - t$, and GP profit, $p + t - c$, which gives $b(s) - c$. Therefore, the difference in utilitarian welfare between being treated by a specialist relative to a GP is again given by $B(s) - C(s) - (b(s) - c)$. Note that this definition of utilitarian welfare is independent of any profit providers may have: this is because any loss of consumer surplus (patient benefit net of transfer) is exactly offset by the increase in profit for the provider. Alternative definitions of welfare are possible. For example, within the economics of regulation, welfare could be defined such that consumer surplus has a higher weight than provider profits, so that leaving positive profits to the provider reduces welfare. However, within this set-up, a regulator (for example, a public or private insurer) could always choose a payment mechanism that leaves zero profits to both the GP and the specialist, leaving the utilitarian welfare unchanged. This is because in our model the regulator could always change the lump-sum payments t and T to ensure the profit is zero, for example by lowering t when p is increased. Provider profits would enter (negatively) the welfare function only when profits are positive and the regulator gives a higher weight to patient net surplus than to profits (Baron and Myerson, 1982).

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 characterise 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 Eqs. (14), (15), (21), and (22) we obtain:

$$\begin{aligned}\Delta\mathcal{B}(\overline{s},\overline{s}) &= \overline{\Lambda}_H\left(\overline{\delta}_H-\overline{\delta}_L\right)\left[B(\overline{s})-b(\overline{s})\right]+\underline{\Lambda}_H\left(\underline{\delta}_H-\underline{\delta}_L\right)b(\underline{s})(1-\omega)>0,\\ \Delta\mathcal{B}(\overline{s},all) &= \overline{\Lambda}_H\left(\overline{\delta}_H-\overline{\delta}_L\right)\left[B(\overline{s})-b(\overline{s})\right]-\underline{\Lambda}_H\left(\underline{\delta}_H-\underline{\delta}_L\right)\left[B(\underline{s})-b(\underline{s})\right]\gtrless 0,\\ \Delta\mathcal{B}(all,all) &= \Delta\mathcal{B}(all,\overline{s})=0.\end{aligned} \tag{26}$$

The health gradients are the direct result of inequalities in specialist treatments in the four scenarios.[19] 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. $\Delta\mathcal{B}(\overline{s},\overline{s})>\Delta\mathcal{B}(\overline{s},all)$.[20]

We consider the introduction of two policies. The first policy relates to tightening the access to specialist services and induces specialists to treat only high-severity patients. For example, this could involve lowering the DRG reimbursement tariff. The second policy relates to tightening the referral system, which induces GPs to refer only high-severity patients. This could involve increasing the fee paid to the GPs. We discuss these policy interventions in turn in Propositions 5 and 6.

Proposition 5. *Consider a policy that tightens specialist treatment by inducing specialists to treat only high-severity patients, as opposed to all patients. The policy increases utilitarian welfare if $-\Delta N\,B(\underline{s})>b(\underline{s})(1-\omega)$. 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 allocative efficiency but does not affect health inequities.*

The policy of inducing specialists to treat only high-severity patients involves two possible transitions.[21] Consider a health system where the GPs refer only high-severity patients, but the specialists have an incentive to treat all patients, which corresponds to scenario (3) in Fig. 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). This policy again increases allocative efficiency but also increases health inequities. An equity-efficiency trade-off arises. Instead, inducing the specialist to treat only high-severity patients, a move from scenario (4) to (2), increases allocative efficiency, but does not affect health inequities. For these results to hold, the condition $-\Delta N\,B(\underline{s})>b(\underline{s})(1-\omega)$ 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 $\Delta N\,B(\underline{s})<0$), 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 6. *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. Incentivising GPs to refer only high-severity patients increases utilitarian welfare if $b(\underline{s})(1-\omega)>\frac{\overline{\lambda}_H\left(1-\overline{\delta}_H\right)+\overline{\lambda}_L\left(1-\overline{\delta}_L\right)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta N\,B(\overline{s})$, and an equity-efficiency trade-off arises. (ii) Suppose the specialist treats all patients. Incentivising GPs to refer only high-severity patients increases utilitarian welfare if $-\Delta N\,B(\underline{s})>\frac{\overline{\lambda}_H\left(1-\overline{\delta}_H\right)+\overline{\lambda}_L\left(1-\overline{\delta}_L\right)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta N\,B(\overline{s})$, and again an equity-efficiency trade-off arises.*

The policy of incentivising GPs to refer only high-severity patients also involves two possible transitions, depending on whether the specialist treats only high-severity patients or all types of patient. 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 $\Delta W(\overline{s},\overline{s})>\Delta W(all,\overline{s})$ or $b(\underline{s})(1-\omega)>\frac{\overline{\lambda}_H\left(1-\overline{\delta}_H\right)+\overline{\lambda}_L\left(1-\overline{\delta}_L\right)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta N\,B(\overline{s})$. Given

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

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

[21] 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 $-\Delta N\,B(\underline{s})>b(\underline{s})(1-\omega)$.

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 makes in treating the high-severity patients. If this condition holds, then incentivising GPs to refer only high-severity patients 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, in which specialists have incentives to treat all patients, which is described under scenario (4). Then, incentivising GPs to refer only high-severity patients, a move from scenario (4) to (3), increases welfare if $\Delta W(\overline{s}, all) > \Delta W(all, all)$ or $-\Delta NB(\underline{s}) > \frac{\overline{\lambda}_H\left(1-\overline{\delta}_H\right)+\overline{\lambda}_L\left(1-\overline{\delta}_L\right)}{\underline{\lambda}_H\underline{\delta}_H+\underline{\lambda}_L\underline{\delta}_L}\Delta NB(\overline{s})$. This condition requires that the welfare loss of those high-severity patients for which the 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 increases utilitarian welfare, which involves incentivising GPs to refer only high-severity patients, it also increases inequities in access to specialist services, 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. One policy is to tighten the gatekeeping role of GPs to induce them to refer selectively only the more severe patients to hospital specialists. A second policy is to restrict 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 GPs.

This study has provided a theoretical framework to assess these policy interventions and has investigated whether such policies generate a tension between the ubiquitous policy objective of reducing inequalities in access and health, and improving the allocative efficiency of health systems, which we measure in terms of utilitarian welfare. 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 hospital specialist is financed through a DRG-based reimbursement system. Patients differ in severity of condition and in socioeconomic status, and the GP receives an informative signal on the severity of the patient following an examination. The signal 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 inequities in access to specialist services and in health are higher in health systems where GPs refer more selectively only high-severity patients. Instead, inequities are smaller in health systems where GPs refer all patients. We show that policies that induce GPs to refer more selectively, for example by increasing the fee paid to GPs to treat patients, generally increase utilitarian welfare but also increase inequities in access to specialist services. Policies that induce specialists to treat only high-severity patients increase utilitarian welfare and inequities in access to specialist services when the GP refers only severe patients, but have no effect on 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, cost containment policies are likely to become more prevalent. There is therefore scope for investigating whether these generate a tension with equity objectives within other contexts in the health sector. There is also scope for additional empirical evidence. The empirical literature has 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 aimed at reducing referrals and restricting 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.

CRediT authorship contribution statement

Oddvar Kaarboe: Concept, Design, Formal analysis, Writing – original draft, Writing – review & editing. **Luigi Siciliani:** Concept, Design, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proofs and supplementary calculations

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jhealeco.2022.102715.

References

Abdalla, S.M., Yu, S., Galea, S., 2020. Trends in cardiovascular disease prevalence by income level in the United States. JAMA Netw. Open 3, e2018150.

Allard, M., Jelovac, I., Léger, P.T., 2011. Treatment and referral decisions under different physician payment mechanisms. J. Health Econ. 30, 880–893.

Bago d'Uva, T., Jones, A.M., 2009. Health care utilisation in Europe: New evidence from the ECHP. J. Health Econ. 28, 265–279.

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

Baron, D.P., Myerson, R.B., 1982. Regulating a monopolist with unknown costs. Econometrica 91, 1–930.

Baron-Epel, O., Garty, N., Green, M.S., 2007. Inequalities in use of health services among Jews and Arabs in Israel. Health Serv. Res. 42, 1008–1019.

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

Brekke, K.R., Nuscheler, R., Straume, O.R., 2007. Gatekeeping in health care. J. Health Econ. 26, 149–170.

Brekke, K.R., Siciliani, L., Straume, O.R., 2011. Hospital competition and quality with regulated prices. Scand. J. Econ. 113 (2), 444–469.

Cerin, E., Leslie, E., 2008. How socioeconomic status contributes to participation in leisure-time physical activity. Soc. Sci. Med. 66, 2596–2609.

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

Chen, A., Lakdawalla, D., 2019. Healing the poor: The influence of patient socioeconomic status on physician supply responses. J. Health Econ. 64, 43–54.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., Cutler, D., 2016. The association between income and life expectancy in the United States, 2001–2014. JAMA 315, 1750–1766.

Devaux, M., 2015. Income-related inequalities and inequities in health care services utilisation in 18 selected OECD countries. Eur. J. Health Econ. 16, 21–33.

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 Educ. Couns. 58, 265–270.

Ellis, R.P., McGuire, T.G., 1986. Provider behavior under prospective reimbursement: Cost sharing and supply. J. Health Econ. 5 (2), 129–151.

Glazer, A., 2004. Motivating devoted workers. Int. J. Ind. Organ. 22, 427–440.

Glied, S.A., 2000. Managed care. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics, Vol. 1A. North Holland, Amsterdam, pp. 707–753.

Godager, G., Iversen, T., Ma, C.-t.A., 2015. Competition, gatekeeping, and health care access. J. Health Econ. 39, 159–170.

González, P., 2010. Gatekeeping versus direct-access when patient information matters. Health Econ. 19 (6), 730–754.

Greenfield, S., Kaplan, S.H., Ware, J.E., Yano, E.M., Frank, H.J.L., 1988. Patients' participation in medical care. J. Gen. Intern. Med. 3, 448–457.

Griebenow, M., Kifmann, M., 2021. Diagnostics and Treatment: On the Division of Labor Between Primary Care Physicians and Specialists. HcHe Research Paper No. 2021/25.

Harle, A., Buffin, O., Burnham, J., Molassiotis, A., Blackhall, F., Smith, J.A., 2014. The prevalence of cough in lung cancer: its characteristics and predictors. J. Clin. Oncol. 32, 162.

Hashmi, R., Alam, K., Gow, J., March, S., 2021. Prevalence of mental disorders by socioeconomic status in Australia: A cross-sectional epidemiological study. Am. J. Health Promot. 35, 533–542.

Jelovac, I., 2014. Primary care, gatekeeping and incentives. In: Culyer, A.J. (Ed.), Encyclopedia of Health Economics, Vol. 3. Elsevier, New York, pp. 142–145.

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

Mackenbach, J.P., Stirbu, I., Roskam, A.-J.R., Schaap, M.M., Menvielle, G., Leinsalu, M., Kunst, A.E., 2008. Socioeconomic inequalities in health in 22 European countries. N. Engl. J. Med. 358, 2468–2481.

Malcomson, J.M., 2004. Health service gatekeepers. Rand J. Econ. 35, 401–421.

Mariñoso, B.G., Jelovac, I., 2003. GPs' payment contracts and their referral practice. J. Health Econ. 22, 617–635.

National Institute of Mental Health, 2022. Major depression. In: Mental Health Information, Statistics. https://www.nimh.nih.gov/health/statistics/major-depression#part_2567 (Accessed 8 Oct 22).

Starfield, B., 2006. State of the art in research on equity in health. J. Health Polit. Policy Law 31, 11–32.

Street, R., 1991. Information giving in medical consultations: The influence of patients' communicative styles and personal characteristics. Soc. Sci. Med. 32, 541–548.

Tavakoly Sany, S.B., Behzhad, F., Ferns, G., Peyman, N., 2020. Communication skills training for physicians improves health literacy and medical outcomes among patients with hypertension: A randomized controlled trial. BMC Health Serv. Res. 20 (1).

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

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

Verlinde, E., De Laender, N., De Maesschalck, S., Deveugele, M., Willems, S., 2012. The social gradient in doctor-patient communication. Int. J. Equity Health 11 (12).

Waitzkin, H., 1985. Information giving in medical care. J. Health Soc. Behav. 26, 81–101.

Williams, D.R., Mohammed, S.A., Leavell, J., Collins, C., 2010. Race, socioeconomic status, and health: Complexities, ongoing challenges, and research opportunities. Ann. New York Acad. Sci. Biol. Disadvant. 1186, 69–101.