



Gatekeeping in primary care: Analysing GP referral patterns and specialist consultations in the NHS[☆]

Federico Innocenti^a, Barry McCormick^{b,c,d}, Catia Nicodemo^{b,a,e,*},1

^a University of Verona, Italy

^b University of Oxford, United Kingdom

^c CHSEO, United Kingdom

^d IZA, Germany

^e Brunel University of London

ARTICLE INFO

Dataset link: [Gatekeeping General Practice, Specialists, and Health Sector Efficiency \(Reference data\)](#)

JEL classification:

I12

I18

C31

D40

Keywords:

Hospital admissions

Referrals

GPs

Gatekeeping

Panel data

NHS

ABSTRACT

This study investigates the impact of increasing the number of gatekeeper General Practitioners (GPs) on referral rates and specialist treatments. Gatekeeping is a supply-side strategy implemented to control health expenditure and improve efficiency by limiting patient access to services below marginal cost. It aims to address specialist moral hazard by reducing the overuse of expensive diagnostics and replacing them with more cost-effective GP diagnostic information. Using administrative data from 2004 to 2011, we examine whether the availability of gatekeeper GPs in local areas is associated with changes in outpatient referrals and elective admissions. Our findings reveal that increasing GP supply in socioeconomically disadvantaged areas leads to a decrease in both outpatient referrals and elective admissions. However, these effects are less pronounced in prosperous areas or regions with high GP referral rates. Interestingly, we observe that having more GP practices in a specific area implies higher referral rates and elective admissions. These findings offer valuable insights that can assist policymakers in crafting targeted policies to effectively reduce healthcare costs and enhance the overall efficiency of the health system.

1. Introduction

In many European countries (e.g., Italy, Spain, the Netherlands, and the UK), general practitioners (GPs) serve as gatekeepers to specialist care in single-payer national health systems. By managing patient referrals, GPs can regulate access to higher-cost specialist services and enhance efficiency. However, the mechanisms by which gatekeeping improves health sector efficiency have not been fully elaborated.

Evaluations of the US healthcare system's efficiency have acknowledged the potential benefits of supply-side constraints. For instance, [Garber \(2004\)](#) and [Garber and Skinner \(2008\)](#) highlight the absence of such constraints as a factor contributing to the comparatively high cost of healthcare in the US. Meanwhile, gatekeeping is a prominent supply-side restraint in several cost-effective European systems.²

This paper aims to enhance our understanding of how controlling patient access to specialists can improve healthcare efficiency.

[☆] This study received funding from the NHIR and we published a report for the funder. However, the analysis in that report was preliminary and the model was still in development. The published report does not contain the final analysis or fully developed model (Chalkley, Martin, et al. "Elective hospital admissions: Secondary data analysis and modelling with an emphasis on policies to moderate growth". Health Services and Delivery Research 5.7 (2017).)

* Corresponding author.

E-mail address: catia.nicodemo@economics.ox.ac.uk (C. Nicodemo).

¹ Funding Acknowledgement: Nicodemo receives support from by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority European Union's Horizon Europe research and innovation programme. Neither the European Union nor the granting authority can be held responsible for them. The PCR-4-ALL project has received funding under the Horizon Europe research and innovation programme (the grant agreement No 101095606) and from Thames Valley at the Oxford Health NHS Foundation Trust.

² Economic analysis on how gatekeeping improves health sector efficiency focuses on addressing the common issue of patient moral hazard and how GPs can regulate insured patients' access to services priced below their marginal cost ([Marinoso and Jelovac \(2003\)](#), [Malcomson \(2004\)](#), [González \(2010\)](#), [Allard et al. \(2011\)](#), [Brekke et al. \(2007\)](#)).

Our contribution is twofold. First, we develop a model showing that gatekeeping can improve healthcare efficiency by approaching the optimal allocation of diagnostics between GPs and specialists, reducing excessive spending on diagnostics by specialists. When insurers cannot effectively monitor costly diagnostic usage, competition among specialists for patients can lead to unnecessary diagnostic expenses and the underutilisation of GP diagnostic information. In a gatekeeping system, GPs determine which patients to refer to specialists based on a noisy signal of illness severity. When approaching a GP, a patient can have a mild or severe illness. GPs rely on clinical knowledge and can effectively treat mildly ill patients, whereas specialists use more advanced and costly diagnostics necessary to treat severely ill patients. GPs are imperfect insurer agents: they consider income and expected population health gains when deciding upon referrals. We find that GPs estimate the likelihood of severe illness and refer patients to specialist treatment if this likelihood is above a threshold. An increase in the supply of GPs reduces each GP's workload from treatments, increasing such threshold (i.e., reducing referrals by GPs). Instead, fiercer competition among practices reduces the threshold if GPs' per capita fee exceeds their opportunity cost of treating patients. Although each insured patient prefers unlimited access to diagnostics, gatekeeping can lead to higher expected utility from a societal perspective.³

Secondly, we provide empirical evidence that supports our theoretical predictions. We analyse the effects of changes in the GP supply on referrals and specialist treatments using NHS panel data from small areas in England between 2004 and 2011. We use instrumental variables methods (IV) to address potential biases in GP location choice, specifically employing a Bartik instrument (Bartik, 1991). Our findings reveal that increasing the number of GPs in a local area results in fewer specialist appointments and subsequent treatments. This effect is particularly pronounced in deprived areas, where a higher density of GPs decreases both outpatient referrals and elective admissions. Furthermore, areas with smaller practices but the same GP density tend to have more outpatient referrals and elective admissions.

These results are consistent with OECD evidence (Masseria et al., 2009), indicating that Western European health systems with gatekeeping have fewer CAT and MRI scans per population. However, the literature on increasing GP gatekeepers yields mixed findings. While Gulliford (2002) found higher GP supply linked to reduced UK hospital admissions, Harris et al. (2011) found limited evidence of GP supply affecting English emergency admissions. Godager et al. (2015) did not observe a reduction in speciality referrals with increased GPs in Norway, but Fang and Rizzo (2009) found that the impact of GP supply on referrals is influenced by whether GPs are self-employed or working in an HMO.

Our study contributes to this complex landscape by elucidating gatekeeping's role in enhancing healthcare efficiency, particularly in less affluent areas. Nevertheless, we acknowledge that healthcare is a unique sector where reduced usage might not always be desirable. For instance, under-diagnosis can lead to more severe diseases in the long term, potentially incurring higher costs for insurers. Our findings thus depict an interesting trade-off between efficiency and social welfare that policy-makers should address.

The prevailing "altruistic clinician" model suggests physicians consider their interests and patient health when making decisions (Newhouse, 1970; Ellis and McGuire, 1986; Gaynor et al., 2004; Chandra et al., 2012). However, such a model fails to explain gatekeeping, as it

³ An insured patient does not bear the cost of the exact test carried out by a specialist and does not consider her negative externalities on medical performance (e.g., higher waiting time and workload). Therefore, an insured patient always benefits from a referral. However, from the insurer's perspective, it is optimal to save unnecessary test costs and to provide access to an effective health system only to those who need it. Thus, it is optimal to avoid referrals when the likelihood of severe illness is low.

implies that GPs paid per capita would refer all patients, benefiting both patients and GPs. Our model conceptualises GPs as imperfect agents balancing competing priorities — their own interests and maximising population health gains. This aligns with the education sector, where teachers prioritise school-wide well-being over their students' welfare.

The paper is structured as follows: Section 2 describes the model and hypotheses. We present the empirical strategy, data, and results in Sections 3, 4, and 5. Section 6 discusses robustness checks, including whether the findings are similar in less and more deprived areas. Section 7 concludes.

2. A model of GP diagnosis, referrals, and hospital admissions

Consider an economy with a population of homogeneous individuals who may become mildly or severely ill. An individual is ill with probability θ . Illness is severe with probability q and otherwise mild. GPs and specialists can reduce the disutility of illnesses they diagnose and treat. There are two treatments: only specialists can prescribe treatment 1, which gives a gain of z to severely ill patients and a smaller or negative gain \bar{z} to others. Instead, all doctors can prescribe treatment 2, giving all ill patients a gain g . We assume that $z > g > \bar{z}$.

Specialists compete for patients: they maximise patient health gain and provide an exact test, at cost c , for each patient. Test information is verifiable; therefore, specialists treat optimally by prescribing treatment 1 to the severely ill and treatment 2 to the mildly ill.

Instead, GPs receive a noisy signal of patient severity. The corresponding probability that a patient is seriously ill is s . We assume that $s = q + \varepsilon$, where $\varepsilon \sim U[-q, +q]$ is relevant patient information. It follows that $s \in [0, 2q]$. Therefore, we assume that $q \leq 0.5$.⁴ The signal has probability distribution $f(s) = 1/2q$, and distribution function $F(s) = s/2q$. GPs choose a referral threshold s^* , refer a patient to a specialist if $s > s^*$, and give treatment 2 otherwise. Therefore, the probability that a GP does not refer and treat an ill registrant is $F(s^*)$.

Individuals have to decide whether to register with a GP. An insured individual i gets the following utility from registration:

$$U_i = \bar{G}(s^*) - T_i$$

and zero without registration. The function $\bar{G}(s^*)$ is the perceived expected health gain from registration, whereas T_i is the cost of registering and forming a relationship with a GP.⁵ $\bar{G}(s^*)$ decreases with s^* because at a higher s^* fewer registered patients are referred and treated optimally. Individual i registers if $U_i > 0$. There is a mass of patients with different costs T_i . We define $N(s^*)$ as the mass of patients that choose to register. $N(s^*)$ is decreasing and concave in s^* , following the properties of $\bar{G}(s^*)$.

Each GP receives a per-registrant fee of r . Let λ be the inverse of the number of GPs chosen by the insurer per practice. Thus, $\lambda N(s^*)$ is the number of registrants per GP, of whom $\theta F(s^*) \lambda N(s^*)$ become sick and are treated by the GP. A GP costlessly diagnoses and refers but requires k hours per patient to give treatment 2. Hence, $k \theta F(s^*) \lambda N(s^*)$ hours are spent treating patients. This activity is associated with an opportunity cost of w per hour. The GP is an imperfect agent of the insurer – the NHS in the UK – and chooses s^* to maximise a combination of the expected net health gain of her registered patients $G(s^*)$, with weight α , and her private utility:

$$V(s^*) = \lambda N(s^*) [r - wk \theta F(s^*)] + \alpha G(s^*) \tag{1}$$

GP's private utility comprises per capita income and the opportunity cost of treating patients she has chosen not to refer. An increase in s^* (fewer referrals): (i) increases GP work hours by increasing the patients

⁴ A property of this specification is that GPs are unwilling to estimate a high probability of severe illness if the illness is rare ($q \rightarrow 0$).

⁵ $\bar{G}(s^*)$ differs from the actual expected health gain $G(s^*)$ – that we introduce below — because the patient is insured, and thus disregards the cost of care.

that she treats, $\theta F(s^*)$, and (ii) reduces the sick patient's benefit from the health system, and hence the number of registrants, $N(s^*)$, and fee income, $r\lambda N(s^*)$.

The timing of the model is the following:

1. The insurer chooses the supply of GPs λ and a GP fee per patient r .
2. GPs maximise $V(s^*)$ by choosing a threshold level of estimated severity above which she refers, and below which she treats patients.
3. Patients choose whether to register with a GP and be able to receive care if ill.
4. Individual illness is revealed, and ill registrants contact their GP, who receives an imperfect signal of the probability that the patient is severely ill and may refer.

We proceed by backward induction. Stages 3 and 4 are straightforward. Stage 1 is not modelled. Therefore, in the following, we focus on stage 2.

2.1. Expected net health gain

The expression for the expected health gain net of treatment cost, $G(s^*)$, has three parts. The first term is the expected net health gain from GP treatment for patients with $s < s^*$. The second term gives the expected net gain for mildly ill referred patients diagnosed at cost c . The third term gives the net health gain for those referred, diagnosed to be severely ill, and given treatment 1. Thus, $G(s^*)$ is given by:

$$G(s^*) = \theta \left\{ \underbrace{g \int_0^{s^*} f(s) ds}_{\text{Net Gain from GP giving treatment 2 without exact test}} + \underbrace{\int_{s^*}^{2q} (g-c)(1-s)f(s) ds}_{\text{Net Gain from Specialist giving treatment 2 to mildly ill after exact test}} + \underbrace{\int_{s^*}^{2q} (z-c)sf(s) ds}_{\text{Net Gain from Specialist giving treatment 1 to severely ill after exact test}} \right\}$$

This may be simplified to give:

$$G(s^*) = \theta \left[(g-c) + cF(s^*) + (z-g) \int_{s^*}^{2q} sf(s) ds \right] \tag{2}$$

In (2), the expected net health gain per patient has three components: a minimum expected health gain, $\theta(g-c)$, for mildly ill patients given treatment 2 by the specialist, after a test; the saved diagnostic costs, $\theta cF(s^*)$, from patients given treatment 2 by the GP; and the health gain for severely ill referred patients who receive treatment 1, $\theta(z-g) \int_{s^*}^{2q} sf(s) ds$. Increasing s^* has ambiguous effects on $G(s^*)$: it reduces specialist diagnostic costs, but fewer severely ill patients receive the benefits of treatment 1.

Lemma 1. $G(s^*)$ is maximised when the GP referral threshold is $s_e^* = c/(z-g)$.

Proof. The change in $G(s^*)$ from a marginal increase in the referral threshold is given by:

$$\frac{dG}{ds^*} = \frac{\theta}{2q} [c - (z-g)s^*]$$

$G(s^*)$ is increasing if and only if $[c - (z-g)s^*] \geq 0$. Moreover, $G(s^*)$ is concave since $z-g > 0$.⁶ Finally, a sufficient condition for an interior maximum – that is, $0 < s_e^* < 2q$ – is that diagnostic costs are sufficiently low relative to the health gain of specialist treatment, i.e. $c < 2(z-g)q$. \square

The intuition for the concavity of $G(s^*)$ is that if s^* marginally increases from a low level, those no longer referred are unlikely to be severely ill and thus have the most negligible loss of expected health gain from not being precisely diagnosed. Instead, at higher s^* , those

⁶ It follows that $\bar{G}(s^*)$ is also concave in s^* . Indeed, $\bar{G}(s^*)$ corresponds to $G(s^*)$ when setting $c = 0$.

forgoing exact diagnoses are more likely to be severely ill. Thus, the net effect from increasing s^* becomes negative. The concavity of $G(s^*)$ explains selective access to specialists, or “gatekeeping”.

2.2. The GP's choice of the referral threshold

A registrant creates $k\theta F(s^*)$ expected hours of work for the GP. A GP's decision on s^* hinges on the relationship between the opportunity cost of treatment and the per-registrant fee r :

$$r \stackrel{\geq}{\leq} wk\theta F(s^*)$$

Since GPs are constrained by their professional ethic to treat patients less severely ill than s^* , they are usually paid a different hourly wage than other workers. Workers will reject GP posts if the insurer persistently sets r too low. Therefore, we assume that, generally, $r \geq wk\theta F(s^*)$.

Proposition 1. GP's referral threshold s^* is less than s_e^* , the threshold that maximises the expected health gain of registrants if the GPs implicit hourly wage is weakly higher than w . In this case, GPs refer more patients than is efficient.

Proof. The GP's maximisation problem is:

$$\max_{s^*} V(s^*)$$

The corresponding first order condition is given by:

$$\frac{dV}{ds^*} = \underbrace{\lambda N'(s^*)}_{-} [r - \underbrace{wk\theta F(s^*)}_{+}] - \underbrace{wk\theta f(s^*)}_{+} \lambda N(s^*) + \underbrace{\alpha G'(s^*)}_{+/-} = 0$$

Hence

$$\alpha G'(s^*) = wk\theta f(s^*) \lambda N(s^*) - \lambda N'(s^*) [r - wk\theta F(s^*)] \tag{3}$$

Since $r \geq wk\theta F(s^*)$, the right-hand side of (3) is positive. Following the concavity of $G(s^*)$, this implies that it must hold $s^* < s_e^*$.⁷ \square

At the optimum, the GP equates the satisfaction of higher registrant net health gain due to a marginal increase in s^* (fewer referrals), given by $\alpha G'(s^*)$, with two sources of costs to the GP. The first term on the RHS of (3) is the opportunity cost of time due to a GP treating marginally more patients from her list. The second term reflects that choosing a higher s^* reduces the number of registered patients and, thus, GP income.

The GP treats fewer (refers more) patients than what would be efficient because she is an imperfect insurer agent. Although willing to treat some patients, she also values time and registrant per capita revenue, which are reduced by fewer referrals. Because s^* is below the social optimum, more patients register with a GP than if she were a perfect agent for the insurer.

2.3. Comparative statics

We use the Implicit Function Theorem to study how some parameters of the model affect the optimal referral threshold. In particular, we study the effects of increasing (i) GPs supply – assuming that GPs are allocated to existing practices that share patients equally between GPs – and (ii) the elasticity of the demand for registration as a measure of underlying competition.

⁷ The second order condition is

$$V_{ss} = \frac{d^2V}{ds^{*2}} = \lambda N''(s^*) [r - wk\theta F(s^*)] - \frac{wk\theta \lambda N'(s^*)}{q} + \alpha G''(s^*) < 0$$

If $r \geq wk\theta F(s^*)$, then the first term is negative since $N'' < 0$. From the expression for G'' , a sufficient condition for $\frac{d^2V}{ds^{*2}} < 0$ is that $\alpha(z-g) > -\frac{wk\theta \lambda N'(s^*)}{q}$.

GPs supply.

$$\frac{ds^*}{d\lambda} = \underbrace{\{wk\theta f(s^*)N(s^*) - N'(s^*)[r - wk\theta F(s^*)]\}}_{+} / \underbrace{V_{ss}}_{-} < 0 \quad (4)$$

Referrals by a GP reflect a tension between self-interest, which implies referring all sick patients, and the professional ethic to maximise expected patient health gain net of cost, and hence to refer only those with $s > s^*$. An increase in the supply of GPs (smaller λ) reduces registrants per GP, $\lambda N(s)$. In Eq. (1), λ acts as a scaling factor of the costs to a GP of reducing referrals: fewer registrants per GP leads to lower marginal costs of increasing s^* and reducing referrals. Indeed, if the GP has to treat a smaller fraction of patients, (a) the GP needs less extra time – that is, $k\theta f(s^*)N(s^*)$ – to treat them when increasing s^* , and (b) the number of quitting patients – that is, $N'(s^*)$ – and lost per capita income – that is $wk\theta F(s^*)$ – is also lower. Thus, an increase in the supply of GPs increases the chosen threshold s^* towards the socially optimal level s_e^* .

In the NHS, the per-registrant fee r varies between practices because the fee differs between patients with different socio-economic characteristics. Thus, the difference between the per capita fee r and the opportunity cost of treating patients may vary between areas belonging to different socio-economic categories. The consequence may be to create regional variability in the response of s^* to a change in GP supply. In the empirical analysis, we explore such spatial variation.

Competition.

$$\frac{ds^*}{d\mu} = \lambda \left(\frac{N(s^*)}{s^*} \right) [r - wk\theta F(s^*)] / V_{ss} \quad \text{where } \mu = -\frac{N'(s^*)s^*}{N(s^*)} \geq 0 \quad (5)$$

We analyse competition in GP services as competition between practices at different locations rather than between individual GPs. The effect of competition, as measured by μ , depends on the relationship between GPs per-registrant fee r and the expected opportunity cost of treating patients $k\theta wF(s^*)$. When r exceeds the opportunity cost, increased competition induces the GP to reduce the threshold of referrals s^* and GP treatments. The opposite is true when $r < wk\theta F(s^*)$. Why does increasing competition between practices not always have the conventional expected effect of increasing referrals to raise patient utility? The existence of an outside option w removes part of the threat to income of more competition in the GP market. We explore this by estimating whether having more or less GP practices in a small geographic area, holding constant the supply of GPs per head, influences referrals and elective hospital admissions.

3. Empirical strategy

In this section, we investigate whether, in a healthcare system like the NHS, where gatekeeping plays a central role, there is evidence from extensive administrative data suggesting that increasing the local GP supply can lower the rates of outpatient appointments with specialists and planned treatments within hospitals.

To estimate this, we use a fixed effects panel data model for three types of hospital referrals/admissions – outpatients or elective cases – at the Lower Super Output Area (LSOA) level.⁸ We control for specific characteristics of the areas and primary care variables. The empirical model is as follows:

$$F_{jt} = \beta X_{jt} + \alpha GP_{pt} + \rho GR_{pt} + Otherc_{pt} + z_{jt} + d_{jt} + \omega_{pt-1} + \sigma_j + \mu_t + \epsilon_{jt}$$

⁸ Lower Layer Super Output Area (LSOA) is a geographic area. Lower Layer Super Output Areas are a geographic hierarchy designed to improve the statistical reporting of small areas in England and Wales. Lower Layer Super Output Areas are built from groups of contiguous Output Areas and have been automatically generated to be as consistent in population size as possible. They typically contain from four to six Output Areas. The Minimum population is 1000, and the mean is about 1500.

where F_{jt} represents the number of outpatient referrals or hospital admissions (elective) per 1000 residents at each LSOA (j) in each year t ; X_{jt} is a vector of socio-economic characteristics that are time-varying at LSOA j in time t – which includes a percentage of gender, age, and ethnicity; and β is a vector of the slope effects of these variables.

The key explanatory variables that capture the supply of GP services in each year t are a measure of both (i) the number of GPs employed and (ii) Full Time Equivalent (FTE) GP employment. While the latter may be a more representative measure of local labour supply. These variables are calculated at the Primary Care Trust (PCT) level for two reasons: first, there are some LSOAs without general practices, and second, the PCT was the area of reference for the GP policies.⁹ The term GP_{pt} represents either the density of GPs per 1000 population in the PCT p in time t , or the GPs FTE supply in PCT p in time t . The variable GR_{pt} represents the density of general practices at time t in PCT p per 1000 of population, and captures whether GPs in an area are concentrated into a few practices.

We also control for other new primary care delivery approaches: Walk-in-Centres (WIC) and Out-of-Hours services (OOH), which are labelled 'other cost prescribing centres' ($Otherc_{pt}$). In each case, we calculate the density of these providers by dividing the number of each type of provider by the local PCT population in 1,000 s . Since the geographic area of PCTs does not change over time, the fixed LSOA effects control for variation in PCT size, allowing the density variables to capture the average effects of within-LSOA changes in the density of traditional practices, or Walk-in-Centres.

The terms z_{jt} and d_{jt} capture measures of the urban and deprivation local area dummy variables. Finally, we control for within year effects (μ_t), prevalent diseases in the year before (ω_{pt-1}), and LSOA (σ_j) fixed effects.¹⁰ ϵ_{jt} is a normally distributed random error term. The standard errors are clustered at the PCT level.

The link between a rise in GP supply and a decrease in referrals might be influenced by the sub-specialities of GPs. For instance, newly entered GPs with more recent training might possess enhanced speciality skills, improving diagnostic abilities and reducing referral necessity. However, it is important to note that general practice is already a specialised field of medicine in the UK. GPs undergo standardised training through dedicated primary care programs, which should reduce variability in diagnostic and referral skills.

3.1. Identification and instrumental variables

There are several issues to consider when estimating the model. Firstly, the number of headcount GPs and FTE GPs in each Primary Care Trust (PCT) may not be exogenous and could be correlated with unobserved factors driving higher healthcare demand. Consequently, an increased GP supply is likely linked to unseen demand influences that raise hospital admissions, even after accounting for constant differences between areas and socio-economic factors.

On the other hand, there might be a negative bias in the estimates of GP effects. GP location decisions are influenced by the GP remuneration system, with certain types of payment linked to the mix of patient types, which can vary across areas. For instance, payments can be associated with patient age, deprivation levels, fee-per-item payments for services like night visits for high-risk groups, and incentives for meeting quality targets. As a result, areas with varying health statuses could experience

⁹ PCTs hold the budget and provide the health care organisation for geographical areas with about 400,000 residents. A Primary Care Trust (PCT) contains, on average, 210 LSOAs.

¹⁰ To avoid any possibility of the endogenous recording of conditions following hospital admission, we use the prevalence data for the year prior to that for the year of study for hospital admissions. Including these effects allows us to identify the impact of variations in primary care supply on hospital admission at the LSOA area level.

different rewards per patient. Consequently, the association between GP supply and hospital admission, as well as the impact of GP supply on it, could be either positive or negative.

To address the potential endogeneity issue, we use two instruments for each GP supply variable, representing both headcount numbers and FTE (full-time equivalent). For headcount GPs, we adopt the methodology proposed by Altonji and Card (1991), which takes into account the significance of immigrant enclaves. We use instruments based on recent flows of country-specific immigration to the United States and the distribution of past migrants' country-specific destinations. This approach relies on the observation that immigrants tend to cluster in cities where prior immigrants from their country of origin have already settled. By leveraging these city-specific factors, the "network" instrument achieves identification. Altonji and Card (1991), Card (2001), and Card and Lewis (2007) have successfully employed this instrument to estimate the causal effect of immigration on the labour market outcomes of U.S. natives.

This approach typically relies on an instrumental variable that assigns different numbers of immigrants to each city each year without influencing labour market outcomes in the city through any channel other than its impact on immigration flows. We can use the same instrument for our variable headcount GPs in this context. The number of GPs located in a PCT over time is instrumented by the share of GPs located in this area in 1980 multiplied by the total number of GPs in year t . This year was selected as it provides a sufficiently historical measure to mitigate endogeneity concerns while still relevant to subsequent healthcare development patterns. In addition, The 1990s saw a significant influx of overseas GPs into the UK healthcare system. These international medical graduates tended to gravitate towards areas with established networks of their compatriots. Using the distribution of GPs from the 1980s as our instrument, we aim to mitigate potential endogeneity arising from this network-driven migration pattern. This approach helps us capture a GP distribution that predates the substantial overseas influx, thereby providing a more exogenous measure of GP supply that is less influenced by the subsequent migration trends of the 1990s and beyond.

Specifically, let GP_{pt} be the total population of GPs resident in the PCT p in year t , and SGP_{p1980} the share of that GPs resident in PCT p in year 1980. The share of GPs in 1980 in area p is calculated as:

$$SGP_{p1980} = \frac{GP_{p1980}}{\sum GP_{1980}}$$

We then construct the imputed stock of GPs supply in PCT p in year t as follows:

$$\widehat{GP}_{pt} = SGP_{p1980} * \sum GP_t$$

We use this to forecast the supply of GPs in PCT p in the year t as the instrument for the explanatory variable GP_{jt} in the hospital admissions equations.

For the FTE GPs, we also use another instrument to take endogeneity into account. We use the shares of female GPs per 1000 of the population in PCT p in time t as instruments. We observe that the percentage of time allocation by gender is different: women prefer to work fewer hours than men (HSCIC, 2012). The last data published from the Health and Social Care Information Centre show that in 2015, 54.4% of GPs were female, an increase from 52.4% in 2014. The number of female GPs has steadily increased since 2005, when the proportion was 42.5%. However, if one considers only FTE GPs, women account for just under half of the GPs working in England—in 2015 49.1% of FTE GPs were female. This instrument should be correlated with the work hours but not the number of hospital admissions. Male and female GPs may exhibit different referral patterns, possibly due to varying risk tolerances or communication styles. Hospital-level decisions are based on clinical need, not referring to GP characteristics, and ethical guidelines prohibit discrimination based on GP gender. The female GP

proportion likely affects utilisation through its impact on GP supply and accessibility rather than direct treatment effects.

Our analysis is performed by a two-stage least square model (2SLS), in which we correct the standard errors to control for heteroscedasticity. Evidence for GP supply using both headcount and FTE data is provided to give a comprehensive account.

4. Institutional setting and data

4.1. NHS UK system

The UK's National Health Service (NHS) operates a healthcare system where patients must register with a General Practitioner (GP) to access most medical services. General practices (GP practices) are the cornerstone of primary healthcare in many countries, including the UK. Central to their operation is the patient registration process, where individuals choose and enroll with a specific practice based on proximity or preference. Once registered, each patient is assigned a dedicated General Practitioner (GP) who acts as their primary healthcare provider. This arrangement ensures continuity of care as the GP becomes familiar with the patient's medical history, ongoing health concerns, and treatment preferences. Patients cannot directly visit a specialist without a referral from their GP, making the GP's role crucial in the healthcare pathway. When a GP decides a patient needs specialist care, they usually refer the patient to a hospital rather than to a specific specialist. This is because deciding which specialist will see the patient often depends on factors within the hospital, such as the availability of specialists and current waiting lists. The GP typically refers the patient to a relevant hospital department or speciality rather than an individual doctor. For example, they might refer a patient with heart problems to the cardiology department or someone with joint pain to the rheumatology department. Patients generally choose which hospital to visit for their specialist appointment. Most people choose hospitals near their homes or workplaces for convenience. However, patients may sometimes choose a hospital that is further away if it has shorter waiting times or a particular reputation for treating their condition. The hospital will then review the referral and assign the patient to an appropriate specialist based on the GP's information and their internal processes. The patient will then be contacted with details of their appointment. This system allows for more flexibility in managing patient care and hospital resources. However, it implies that patients do not usually know which specialist they will see until their appointment is scheduled.

Importantly, the NHS provides healthcare services that are free at the point of use and funded through taxation, ensuring that financial barriers do not prevent access to care. Patients who are not formally registered with a GP can still access emergency services and some community health services. However, they are not considered part of the GP's patient list for routine care and analysis purposes. In the UK, very few patients are not registered with GPs due to the limited prevalence of private healthcare services. More information about the structure of the NHS can be found on the NHS website (<https://www.england.nhs.uk/>). Finally, GPs are hired by each practice but are subject to a national contract. The latter defines the working conditions for all GPs. Therefore, there cannot be differences in GPs' workload. This characteristic of the NHS justifies our assumption in the model that, in each local practice, GPs share patients equally.

4.2. Clinical data

The *Hospital Episode Statistics* (HES) provide information concerning all inpatients and outpatients admitted to NHS hospitals from 1989-90 onwards. It includes private patients treated in NHS hospitals, patients resident outside of England, and care delivered by Treatment Centres (including those in the independent sector) funded by the NHS. Each patient record contains detailed information, including clinical

information, patient characteristics, such as age and gender, administrative and location information, such as the admission method, and the geography of treatment and residence. Since our focus is on GP influence on admissions, our analysis concerns only the 'first admission' to the hospital, which the GP is most likely to influence, rather than admissions for continuing treatments.

It is crucial to control for the prevalence of diseases to explain referrals and admissions. The Quality and Outcomes Framework (QOF) provides valuable clinical information concerning the prevalence of twenty-two specific diseases in 2013.¹¹ In our study, we consider just eleven clinical conditions, which are the clinical domains set up for the year 2004 and available in all the years of the study. The specific conditions used are: Coronary Heart Disease, Left Ventricular Dysfunction, Stroke and Transient Ischaemic Attack, Hypertension, Diabetes Mellitus, Chronic Obstructive Pulmonary Disease, Epilepsy, Hypothyroidism, Cancer, Mental Health, and Asthma. These are also the most frequent illnesses that influence the demand for hospital admissions.

4.3. Lower super output area and deprivation controls

Anonymous patient records were extracted by financial year (from 1 April to 31 March) and aggregated at LSOA. In 2011, 32,482 LSOAs were established in England. We use ONS mid-year population estimates to calculate LSOA populations, and these data are linked to those of individual characteristics at the LSOA level — such as the percentage by gender, ethnicity, and age. Small areas are mapped to 151 PCTs locked at 2011 boundary configurations. From 1 October 2006, 303 PCTs merged into 151 PCTs, now Clinical Commissioning Groups (CCGs). Socio-economic status at the LSOA level is measured using the deprivation domain of the English Indices of Deprivation (Noble et al., 2008).¹²

4.4. The supply of GPs, the size of practices, and hospital admissions

It is critical for this study to measure the supply of GP services carefully. For this purpose, we link the HES data with the 'General Medical Practices Exeter Payments' data and the 'Practitioners of NHS Connecting for Health' data for the period 2004–2011. The Exeter data concern current GPs in traditional GP practices and give both headcount and FTE information. The NHS Connecting for Health data include all prescribing GPs – including employees of non-traditional providers such as WICs – but gives only headcount and not hours worked (FTE) information. Therefore, our study combines FTE and headcount data to check the robustness of the results to alternative data sources.

To recognise both traditional and new primary care delivery models, practice density at the PCT level is measured using two variables: (i) the number of traditional GP practices per 1000 population (density of practices), and (ii) the total number of Walk-in-Centres (WIC) and Out-of-Hours services (OOH) per 1000 population. The Information Centre supplies the data for both series as part of the information on Prescribing Centres.

There were about 10,100 general practices in the UK in 2011, and patients wishing to receive NHS primary care must register with a

¹¹ Prevalence data are used within QOF to calculate points and payments within each clinical domain area. The Quality and Outcomes Framework (QOF) is a system that remunerates general practices for providing good quality care to patients and helps fund work to improve the quality of health care delivered further. It is a fundamental part of the General Medical Services (GMS) Contract, introduced on 1st April 2004.

¹² The deprivation index combines information regarding the proportion of individuals living in low-income households with indices of crime, education, employment rates, health status, and environmental quality. We exclude the health component of the deprivation index to avoid potential endogeneity but use controls for ten levels of the adjusted deprivation index.

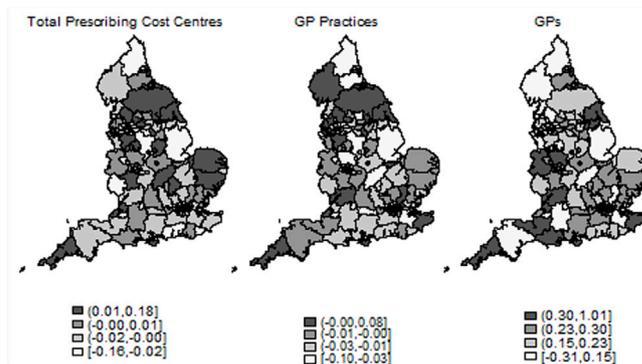


Fig. 1. Distribution of general practices and GPs per 100k pop at PCT in 2011.

single practice. In 2011, there were approximately 50 GP practices in a typical PCT of 300,000 persons, with PCTs on average having 4.3 FTE GPs per practice. The mean number of GPs per practice varies considerably between PCTs, with some as low as three and others as high as seven. GPs may be partners who share ownership of a practice or salaried in receipt of a wage for a specified number of sessions/hours. The payments from the NHS to GPs are made to the practice and not the individual. The funding is a formula based on the population need characteristics of the patients on each GP list and is independent of the number of elective or emergency admissions that patients incur or the treatments given by GPs. Since 2004, GPs have been able to choose whether to provide 24-hour care or transfer responsibility for out-of-hours services (OOH) to PCTs, now Clinical Commissioning Groups (CCGs).

4.5. Summary statistics

Tables 4–6 in the appendix provide the description of the variables, the summary description of the variables used to study the period 2004–2011, and the summary statistics of all periods. Both outpatient referrals and elective admissions increase over time, with referrals increasing most rapidly. The mean proportion of the elderly in local populations increases slightly for both males and females. The mean PCT density of (i) GPs per head increased by about 27% between 2004 and 2011, and (ii) GP FTEs increased by about 11% between 2001–2009 but then declined by 3% in the period 2009–2011. The mean number of traditional GP practices per 1,000 population is unchanged at one practice per 5,000 persons. Mean age and ethnic population proportions have both increased. The WICs and OOHs are primary care services that started around 2005–2007 and, as reported by Exeter data, remain few in number. Fig. 1 illustrates the distribution of GPs, GP practices, and Out-of-Hours Prescribing Centres across PCTs in 2011. This visualisation reveals notable variations in healthcare service provision among different areas. The figure highlights disparities in the availability of these three essential healthcare resources, offering insights into the diverse healthcare landscapes across PCTs during that year.

5. Results

For both referrals and elective admissions, two sets of results are presented corresponding to the two GP supply variables per head of the PCT population: (i) total GP headcount and (ii) FTE (hours) data. Each set of results also gives estimates of the influence of the density of practices, which is measured using the number of GP traditional practices per head of population and the density of other 'prescribing cost centres,' which include the introduction of 'Walk-in Centres' and 'Out of Hours' Service Centres. We provide estimates for both OLS and

Table 1
Estimates of models of GP referrals to specialists per 1000 population.

	Headcounts GPs				FTE GPs			
	OLS		2SLS		OLS		2SLS	
	M1	M2	M3	M4	M1	M2	M3	M4
GPs Density PCT 1k pop	11.63 (21.521)	-3.762 (20.977)	72.335 (71.025)	64.501 (105.226)	3.799 (29.373)	-18.808 (32.947)	-35.766 (31.222)	-74.863** (36.48)
G.Practices Density PCT 1k pop		299.142*** (108.423)		82.005 (335.012)		329.321** (143.107)		454.936*** (125.128)
WIC-OOH Density PCT 1k pop		1151.910* (683.615)		1206.622 (764.381)		1163.573* (683.954)		1189.345* (693.532)
N	298,801	298,801	298,801	298,801	298,801	298,801	298,801	298,801

Data on Referrals are drawn from the Hospital Episodes Statistics. The following variables are also included: prevalence of local area diseases from QoF, years, rural indicator, and Index of deprivation in the LSOA decile fixed effects. Standard errors in parenthesis. Robust and clustered standard errors at PCT level.

* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

Table 2
Estimates of models of elective hospital admissions per 1000 population: alternative method of GP measurement.

	Headcounts GPs				FTE GPs			
	OLS		2SLS		OLS		2SLS	
	M1	M2	M3	M4	M1	M2	M3	M4
GPs Density PCT 1k pop	-21.556*** (4.194)	-26.023*** (3.829)	-6.598 (4.724)	-10.317 (10.026)	-0.625 (3.752)	-0.99 (4.206)	-7.985** (3.766)	-9.980** (4.598)
G.Practices Density PCT 1k pop		83.178*** (28.501)		33.218 (53.421)		2.621 (27.519)		22.765 (30.416)
WIC-OOH Density PCT 1k pop		87.711 (57.892)		100.299 (63.092)		109.023* (65.746)		113.156* (63.939)
N	298,801	298,801	298,801	298,801	298,801	298,801	298,801	298,801

Data on Elective Hospital Admissions are drawn from the Hospital Episodes Statistics. The following variables are also included: prevalence of local area diseases from QoF, years, rural indicator, and Index of deprivation in the LSOA decile fixed effects. Standard errors in parenthesis. Robust and clustered standard errors at PCT level.

* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

2SLS; the latter allows GP density to be endogenous and estimated using instrumental variables methods, as described in Section 3.

Tables 1 and 2 show the results for the estimation of referrals and elective admissions, respectively. Our preferred specification is Model M4, which incorporates a comprehensive set of control variables and addresses potential endogeneity concerns. Additionally, given that the NHS workforce is typically measured in Full-Time Equivalents (FTE) rather than headcounts, we prioritise the estimations using FTE GPs. This approach aligns more closely with standard NHS practices and accurately represents GP availability and its impact on healthcare usage.

The influence of an increase in GP density on elective admissions is consistently negative across the models (M1-M4) for both headcount and FTE data (Table 2). This consistency is robust to alternative specifications for the prevalence of local area disease. Overall, allowing for the endogenous choice of GP location increases this measured effect, but it also holds in OLS estimates. The estimated parameter is larger in the headcount data than the FTE data but significant in both. Consider a GP with 1,200 patients who can expect to undergo 177 elective admissions each year, including day cases, but not maternity admissions. The 2SLS FTE results estimate that adding one extra FTE GP to the practice may reduce this by 10–15 elective admissions per annum. Thus, although statistically significant, the effect alone could not justify one extra FTE GP. Nevertheless, since GPs provide services – in addition to reducing elective admissions – this cost reduction contributes to an overall case of hiring one extra FTE GP.

It is instructive to consider the estimates for first outpatient referrals to confirm whether an increase in the supply of GPs also reduces referrals, as would be expected given our model. Estimates using the FTE data give a highly significant relationship in which increased GP supply reduces outpatient referrals. The slope effect is substantial. The

FTE data imply that the small practice with a patient list of 1,200 would, on average, expect about 493 first outpatient appointments each year, and a 0.2 FTE increase in GP services may reduce this level of outpatient referrals by about 16 p.a. or a little over one per month. Thus, the FTE data overall suggest that an extra 0.2 FTE GPs may reduce referrals by 16 p.a. and ensuing elective admissions by 2-3 p.a.

Tables 1 and 2 also show the effects of practice density on referrals and admissions. In both the NHS headcount and FTE data, increasing the number of practices, with GPs per head of population constant, has a consistently positive effect on first outpatient referrals and elective admissions. A 10% decrease in practice density (increase in practice size) is found to reduce (i) practice referrals by 2.2% from the sample mean of 344 referrals per annum per 1000 patients and (ii) elective treatments by 0.50%. These effects hold consistently in both datasets and estimation methods. This may be interpreted as gatekeeping being less of a constraint on patient demand in areas with many smaller practices and, thus, more significant competitive pressures from patients.

The relationship between GP supply and healthcare utilisation exhibits complexity, as evidenced by variations across our different model specifications. The discrepancies observed between OLS and 2SLS results suggest the presence of endogeneity, which our instrumental variable approach aims to address. The differences between headcount and FTE measures highlight the importance of considering how GP availability is quantified. FTE measures may better capture actual service capacity, while headcounts might reflect broader access points. These inconsistencies underscore the multifaceted nature of primary care provision and its impacts.

Finally, we explore a potential difference between areas the modelling has overlooked thus far. In particular, we study the effect of heterogeneous quality of life. The supply of healthcare services may

Table 3
Estimates of models of referrals and elective hospital admissions per 1000 population: Deprived and not deprived areas.

	Deprived areas		Not deprived areas	
	Referrals	Elective	Referrals	Elective
GPs Density PCT 1k pop	-128.532*** (47.594)	-14.817** (5.781)	11.93 (19.299)	-2.685 (5.219)
G.Practices Density PCT 1k pop	573.063*** (133.52)	62.235* (33.586)	173.468 (122.371)	86.525*** (29.399)
WIC-OOH Density PCT 1k pop	896.028*** (246.33)	169.826* (88.887)	278.29 (377.914)	175.192** (84.743)
N	89,762	89,762	86,530	86,530

Data on Referrals and Elective Hospital Admissions are drawn from the Hospital Episodes Statistics. The following variables are also included: prevalence of local area diseases from QoF, years, rural indicator, and Index of deprivation in the LSOA decile fixed effects. Robust and clustered standard errors at PCT level in parenthesis.

* $p < 0.10$.

** $p < 0.05$.

*** $p < 0.01$.

Table 4
Description of variables.

Name of Variable	Description	Source of data
Emergency	Total first admissions per 1k pop at LSOA level in the financial year: emergency	HES
Elective	Total first admissions per 1k pop at LSOA level in the financial year: inpatients	HES
Referrals	Total first admissions per 1k pop at LSOA level in the financial year: outpatients	HES
Female pop (%)	Percentage of female population at LSOA level	ONS
Female population +60 (%)	Percentage of female population aged 60+ at LSOA level	ONS
Male pop. +65 (%)	Percentage of male population aged 65+ at LSOA level	ONS
Black ethnicity (%)	Percentage of black ethnicity at PCT level	ONS
Asian ethnicity (%)	Percentage of Asian pop at PCT level	ONS
Headcount GPs Density at PCT 1kpop	Number of GPs per 1k pop at PCT level	ODS
FTE GPs Density at PCT 1kpop	Number of GPs FTE per 1k pop at PCT level	HSCIC
G. Practice Density at PCT 1k pop	Number of Gp Practices per 1k pop at PCT level	HSCIC
WIC and OOH Density at PCT 1kpop	WIC and OOH Centres per 1k pop at PCT level	ODS
Revenue per head	NHS Expenditure per capita at PCT level	DH
Deprivation Areas	Index of deprivation at LSOAs in 10 deciles	ONS
Prevalence Diseases	Prevalence of specific diseases per 1k pop at PCT level from QOF	HSCIC

differ between deprived and more prosperous areas of the country, altering how demand is presented and admissions are determined. To better understand how the influence of GP supply may differ according to local deprivation, we compare two groups of LSOAs: the 20% most deprived LSOAs compared to the 20% least deprived LSOAs based on the 2003 Index of Deprivation (IMD) and using the IMD by decile. We apply the previously presented estimation methods for each of the two groups of LSOAs. In Table 3, we report the results. These estimates suggest that increasing GP supply significantly decreases hospital admissions in deprived areas, as we hypothesise, but do not do so in the more prosperous areas. Instead, the effect of fiercer competition across practices (i.e., having more practices holding constant area GP supply) is positive and statistically significant, independent of deprivation. We can draw some conclusions about the effect of deprivation on gatekeeping by using these results and our theoretical predictions — particularly conditions (4) and (5). Since fiercer competition always increases referrals, $r > wk\theta F(s^*)$ must be valid in all areas. In other words, the per capita fee r seems appropriate to cover the opportunity cost of treating patients. This implies that the sign of (5) is always negative. In contrast, the zero effect of increasing GP supply in prosperous areas suggests that condition (4) is close to zero in these areas. The most plausible source of heterogeneity is the probability of being ill θ , which is higher in deprived areas and may drive the positive effect of increasing GP supply.

Our analysis of practice density in non-deprived areas reveals an intriguing pattern: a positive correlation with elective admissions without a corresponding significant increase in referrals. Several mechanisms may explain these seemingly contradictory results. First, areas with higher practice density might facilitate easier access to primary care, leading to earlier detection of conditions requiring elective procedures. Second, the presence of more practices could foster a competitive environment, encouraging GPs to be more proactive in recommending

elective treatments. Additionally, patients in areas with higher practice density might have greater healthcare awareness and be more likely to pursue elective treatments when offered. It is also possible that these areas have developed more efficient pathways for elective admissions that bypass traditional referral processes. These factors could collectively contribute to increased elective admissions without necessarily inflating referral rates, highlighting the complex interplay between primary care structure and healthcare utilisation patterns.

5.1. Robustness checks

To study the relationship between GP supply and the conditional distribution of hospital admissions (outpatient and elective), we use a quantile instrumental variable panel estimator (QIV) developed recently by Chernozhukov and Hansen (2005). The principal identifying assumption of the model is the imposition of conditions that restrict how rank variables (structural errors) may vary across treatment states. These conditions allow instrumental variables to be used to overcome the endogeneity problem and recover the true quantile treatment effects (QTE). The QIV estimator allows us to obtain estimates of the influence of GP supply that vary across the cross-sectional conditional LSOA's hospital admissions distribution. Unlike estimators of the conditional mean, which can be sensitive to values in the tail of the distribution, conditional quantile estimators are inherently more robust to extreme values. In addition, the QIV estimator allows us to address the endogeneity described in the previous section with an instrumental variable identification strategy. To estimate the quantile regression, we use the same variables as above to estimate the 2SLS of the hospital admissions and referrals. Figs. 2 and 3 shows the quantile estimation for the influence of the number of FTE GPs per head of population on referrals and admissions per head. The major variation across the distribution is observed in the outpatient referral equation.

Table 5
Summary statistics.

	2004	2005	2006	2007	2008	2009	2010	2011	Total
Elective per 1k pop	114.92 (35.36)	120.23 (36.70)	124.02 (38.03)	130.64 (37.99)	138.78 (39.75)	142.34 (40.13)	145.19 (40.26)	147.93 (43.36)	133.01 (40.69)
Referrals per 1k pop	261.65 (79.27)	294.42 (86.01)	298.29 (93.87)	321.53 (101.38)	362.30 (132.89)	400.18 (182.78)	405.76 (239.71)	411.86 (312.76)	344.50 (181.26)
Female pop (%)	51.04 (2.30)	51.01 (2.31)	50.98 (2.36)	50.95 (2.48)	50.92 (2.67)	50.90 (2.85)	50.87 (3.11)	50.97 (2.28)	50.96 (2.56)
Female population +60 (%)	11.78 (4.59)	11.83 (4.61)	11.90 (4.65)	12.10 (4.73)	12.28 (4.82)	12.43 (4.88)	12.38 (4.82)	12.56 (4.94)	12.16 (4.77)
Male pop. +65 (%)	6.87 (2.71)	6.92 (2.75)	6.96 (2.80)	7.02 (2.86)	7.14 (2.94)	7.28 (3.04)	7.42 (3.15)	7.27 (3.06)	7.11 (2.92)
Black ethnicity (%)	2.60 (4.40)	2.67 (4.26)	2.72 (4.12)	2.78 (3.99)	2.84 (3.87)	2.88 (3.75)	2.93 (3.64)	2.91 (3.29)	2.79 (3.93)
Asian ethnicity (%)	5.65 (6.89)	5.88 (6.80)	6.11 (6.71)	6.37 (6.66)	6.58 (6.57)	6.78 (6.50)	7.01 (6.46)	7.12 (6.08)	6.44 (6.61)
Revenue per head	1.00 (0.69)	1.00 (0.69)	1.00 (0.68)	1.00 (0.68)	1.00 (0.13)	0.99 (0.13)	1.00 (0.14)	1.00 (0.14)	1.00 (0.49)
Headcount GPs Density at PCT 1kpop	0.83 (0.10)	0.87 (0.10)	0.89 (0.11)	0.93 (0.12)	0.96 (0.13)	1.00 (0.13)	1.04 (0.17)	1.06 (0.18)	0.95 (0.15)
FTE GPs Density at PCT 1kpop	0.61 (0.07)	0.62 (0.07)	0.65 (0.08)	0.64 (0.08)	0.66 (0.08)	0.68 (0.08)	0.68 (0.10)	0.67 (0.09)	0.65 (0.09)
G. Practices Density at PCT 1k pop	0.17 (0.04)	0.17 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)
WIC and OOH Density at PCT 1k pop	0.00 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
N	32482	32482	32482	32482	32482	32482	32482	32482	259856

This includes conventional partnership practices, Walk-in Centres, Out Of Hours Centres, and other Prescribing Cost Centres which include addiction services. Standard Deviation in parenthesis.

Table 6
Summary statistics.

	mean	sd	min	max	p25	p50	p75
Elective per 1k pop	133.01	40.69	1.29	761.50	104.74	129.08	156.99
Referrals per 1k pop	344.50	181.27	2.55	3,623.67	275.79	333.12	155.26
Female pop (%)	50.96	2.56	17.96	206.66	49.84	51.03	400.50
Female population +60 (%)	12.16	4.77	0.04	45.04	8.75	11.85	52.22
Male pop. +65 (%)	7.11	2.92	0.04	35.09	5.00	6.84	15.07
Black ethnicity (%)	2.79	3.93	0.19	22.38	0.74	1.30	8.88
Asian ethnicity (%)	6.44	6.61	0.60	44.00	2.28	3.98	2.49
Revenue per head	1.00	0.49	0.11	2.61	0.69	0.95	0.25
Headcount GPs Density at PCT 1kpop	0.95	0.15	0.54	2.03	0.84	0.94	1.17
FTE GPs Density at PCT 1kpop	0.65	0.09	0.32	1.27	0.59	0.64	1.04
Density G. Practices PCT 1k pop	0.22	0.05	0.11	0.46	0.18	0.21	8.34
WIC and OOH Density at PCT 1k pop	0.01	0.00	0.00	0.03	0.00	0.00	0.01
N			259856				

This includes conventional partnership practices, Walk-in Centres, Out Of Hours Centres, and other Prescribing Cost Centres which include addiction services. Standard Deviation in parenthesis.

The estimates imply that an increase in FTE GPs of the same amount will have a more significant negative effect on specialist referrals in an LSOA where practices have fewer referrals per head than in one with high referral rates. This is consistent with the view that the behaviour of an additional GP is influenced by local practice ‘style’. In areas with a low referral rate, the additional GP reduces referrals by a larger amount, which appears to ‘mirror’ the conservative behaviour of colleagues. We find that the influence of GP supply is unchanging across the distribution for elective admissions except at the top of the distribution. In contrast, its influence on referrals is positive and more prominent in the LSOAs with high levels of referrals, constant for the middle of the distribution and negative at the bottom of the distribution. This estimation gives insight into how the influence of GP supply is more significant for outpatient referrals to hospitals in areas with high referrals, which are the most deprived.

6. Conclusions

This paper discusses the potential benefits of GPs’ gatekeeping in health systems. We explore how gatekeeping can reduce moral hazard among insured patients seeking unnecessary treatments and address issues such as excessive use of costly diagnostics by specialists and

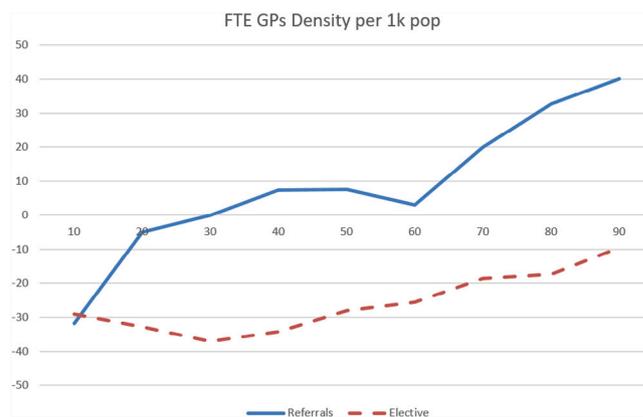


Fig. 2. Quantile regression for GP referrals and hospital admissions: estimated coefficients for FTE GP density.

underutilisation of GP diagnostic information. Our model focuses on two main aspects: (i) understanding the conditions under which GPs

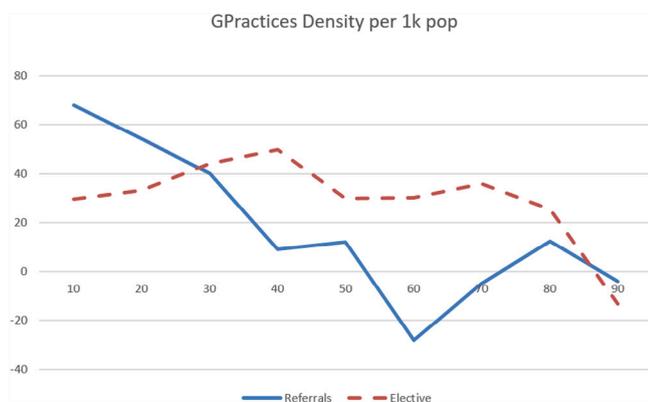


Fig. 3. Quantile regression for GP referrals and hospital admissions: estimated coefficients for GP practice density.

choose to limit access to specialists to mitigate excessive diagnostics, and (ii) examining whether increasing the number of GPs leads to fewer referrals, specialist diagnoses/treatments, and more primary care treatments. The decision to ratio access to specialists can enhance insured individuals' expected net health gain if GPs can effectively identify patients with a lower likelihood of serious illness. The willingness of GPs to restrict referrals depends on their professional ethics, with imperfect agents being more inclined to limit access compared to "altruistic clinicians".

Using NHS administrative panel data, we discovered that an increase in the number of GPs within a specific area leads to a decrease in specialist referrals and a corresponding reduction in planned admissions for their patients. However, this relationship differs between deprived and non-deprived areas. In deprived areas, a higher local GP supply significantly reduces outpatient referrals and elective admissions. These savings account for approximately 20%–30% of the additional GP costs and represent a substantial portion of the case for additional appointments. The variation in outcomes between deprived and non-deprived areas may be attributed to the per capita fee in affluent areas discouraging GPs from expanding their patient lists, thus reducing the incentive provided by more GPs. Alternatively, patient choice in prosperous areas may be more influential, leading to GPs losing patients if they engage in gatekeeping behaviour. Additionally, introducing more practices in a local area increases competition in primary care and results in more referrals and treatments. This implies that policies aiming to enhance the role of general practitioners and reduce demand for specialist services should consider the impact of additional practices.

One limitation is the study's restricted timeframe. Current NHS data policies no longer permit linking these types of datasets or conducting analyses with more recent years, preventing an extension of the investigation to cover a more contemporary period. Furthermore, while using female GP proportion as an instrument provided valuable insights, it is important to acknowledge potential limitations. The relationship between GP gender and practice patterns may evolve over time, and there could be unobserved factors correlated with both female GP proportion and healthcare outcomes. Future research might explore how changing gender dynamics in healthcare affect the validity of this instrument across different periods and healthcare systems.

Future research should consider how efficiency considerations interact with pricing. Pricing below marginal cost induces losses for hospitals. However, it may be efficient if the health gain due to the treatment of poor patients – who otherwise would not access healthcare

– overcompensates for these losses. This paper considers the efficiency gain from restraining access to specialists' healthcare irrespective of income. Suppose GPs' gatekeeping is only partial, and wealthy people have an outside option (i.e., to pay a high price and get specialists' treatment without a GP's referral). In that case, it may be efficient to have different referral thresholds depending on income and incentivise referrals of poor patients.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Any financial interest or personal relationship to be considered.

Data availability

The data that has been used is confidential.

Gatekeeping General Practice, Specialists, and Health Sector Efficiency (Reference data) (Mendeley Data)

References

- Allard, M., Jelovac, I., Leger, P.T., 2011. Treatment and referral decisions under different physician payment mechanisms. *J. Health Econ.* 30, 880–893.
- Altonji, J.G., Card, D., 1991. The effects of immigration on the labor market outcomes of lessskilled natives. In: *Immigration, Trade and the Labor Market*. University of Chicago Press, pp. 201–234.
- Bartik, Timothy J., 1991. Who Benefits from State and Local Economic Development Policies?. *WE Upjohn Institute for Employment Research*.
- Brekke, K., Nuscheler, R., Straume, O., 2007. Gatekeeping in health care. *J. Health Econ.* 26, 149–170.
- Card, D., 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *J. Labor Econ.* 19 (1), 22–64.
- Card, D., Lewis, E.G., 2007. The diffusion of Mexican immigrants during the 1990s: Explanations and impacts. In: *Mexican Immigration to the United States*. University of Chicago Press, pp. 193–228.
- Chandra, A., Cutler, D.M., Song, Z., 2012. *Handbook of Health Economics*, vol. 2.
- Chernozhukov, Victor, Hansen, Christian, 2005. An IV model of quantile treatment effects. *Econometrica* 73 (1), 245–261.
- Ellis, R.P., McGuire, T.G., 1986. Provider behavior under prospective reimbursement: Cost sharing and supply. *J. Health Econ.* 5 (2), 129–151.
- Fang, H., Rizzo, J.A., 2009. Competition and physician-enabled demand: The role of managed care. *J. Econ. Behav. Organ.* 72, 463–474.
- Garber, A.M., 2004. 'Cost effectiveness and evidence evaluation as criteria for coverage policy' health affairs. *Web Exclusive*, May 19.
- Garber, A.M., Skinner, J., 2008. Is American health care uniquely inefficient? *J. Econ. Perspect.* 22 (4), 27–50, 1 2, 4.
- Gaynor, M., Rebitzer, J.B., Taylor, L.J., 2004. Physician incentives in health maintenance organizations. *J. Polit. Econ.* Ma CA.
- Godager, Geir, Iversen, Tor, Ma, Ching-to Albert, 2015. Competition, gatekeeping, and health care access. *J. Health Econ.* 39, 159–170.
- González, Paula, 2010. Gatekeeping versus direct-access when patient information matters. *Health Econ.* 19 (6), 730–754.
- Gulliford, Martin C, 2002. Availability of primary care doctors and population health in England: is there an association? *Journal of Public Health* 24 (4), 252–254.
- Harris, M.J., Patel, B., Bowen, S., 2011. Primary care access and its relationship with emergency department utilisation: an observational, cross-sectional, ecological study. *Br. J. Gen. Pract.*
- Health and Social Care Information Centre, 2012. *Hospital episode statistics, admitted patient care - England, 2011-12*.
- Malcomson, J.M., 2004. Health service gatekeepers. *Rand J. Econ.* 35, 401–421.
- Marinoso, B.G., Jelovac, I., 2003. GPs' payment contracts and their referral practice. *J. Health Econ.* 22, 617–635.
- Masseria, C., Irwin, R., Thomson, S., Gemmill, M., Mossialos, E., 2009. Primary care in Europe European Commission. Access: <http://ec.europa.eu/social/BlobServlet?docId=4739&langId=en>.
- Newhouse, J.P., 1970. Toward a theory of nonprofit institutions: An economic model of a hospital. *Am. Econ. Rev.* 60 (1), 64–74.
- Noble, Michael, McLennan, David, Wilkinson, Kate, Whitworth, Adam, Exley, Sonia, Barnes, Helen, Dibben, Chris, 2008. *The English indices of deprivation 2007*. London: Department for Communities and Local Government.