Will policy to constrain GP referrals damage health? Evidence using practice level NHS emergency admissions administrative data

Abstract

Attempts to control hospital expenditure by managing down General Practitioner (GP) referrals are reoccurring features of UK health policy. However, despite the best efforts of GPs to benchmark referral criteria, patient health may be damaged and other costs created by constraining referrals to targets. This paper adopts an indirect method to indicate whether rationing practice referrals may damage population health by distorting the use of health resources away from patients' interests. We utilise a comprehensive database at practice level that allows us to explore the relationship between referrals and emergency admissions, using a panel fixed effects model of admissions that allows for the endogeneity of referrals. We find that practice referrals are positively and partially correlated with emergency admissions, which is consistent with time-varying practice-level sickness shocks driving the relationship between referrals and emergency care, rather than shocks to the practice willingness to refer, or to system reforms. In this environment, government policy to constrain referrals may make the elective care less responsive to practice-level variations in illness, and thereby lower health.

Keywords: Emergency Hospital Admissions, Referrals, General Practice, England, NHS

JEL: C23, I10, I18

1. INTRODUCTION

English National Health Services (NHS) policy makers have frequently attempted to change general practitioners' (GPs) behaviour as healthcare gatekeepers, as a way of managing the demand for hospital services, and in particular referrals. GPs are able to control patients' access to treatment because they are the major portal through which patients access specialist medical services. This GP referral system helps limit health care expenditure, and in a free-at-delivery system such as the NHS, is key to preventing patients consuming services above the level where patient benefit at least equals cost. A key issue is whether practice referral rates that are high relative to benchmarks can be reduced without harming patient health. However, measurement of health loss due to rationing referrals, when a practice may otherwise exceed benchmarks, is difficult to study since individual patients who are rationed at the margin are not identified.

We adopt an indirect method to study whether rationing practice referrals affects health by exploring the relationship between referrals and emergency hospital admissions. One plausible relationship is that higher referrals enable patients to require fewer emergency admissions. If rationing referrals leads to increased emergency care, this is less healthy for the patient and more costly for the NHS, and may easily offset the savings made by limiting referrals. A second relationship is that practices are subject to 'shocks' over time, which may influence both referrals and emergency admissions. In particular, we consider whether higher GP referrals may be indicating a 'shock' increase in patient illness at a practice, rather than a generous referral policy, by examining the emergency hospital admissions of patients at the practice.

If there is a positive relationship between practice referrals and above-expected emergency admissions, we suggest that this is due to referrals identifying time-varying broad health needs in the local community, whilst a negative relationship could be explained by substitution

between referrals and emergency admissions. We also discuss how a weak relationship might result from shocks to the practice willingness to refer, or system reforms. If referrals are responding to greater local need, then suppressing referrals could have negative consequences for local population health. If this is the case, constraining referral rates will make the system less responsive to health needs, and probably more expensive.

In the last decade, there have been large increases in the average number of referrals and emergency admissions. From 2008/9 to 2017/18, referrals per 1000 population increased by 16% from 210.0 to 243.5 per annum, and emergency admissions from A&E departments per 1000 population by 27% from 63.1 to 80.3 per annum.

Several explanations for the growth in hospital admissions have been proposed. For example, an increase in illness and general frailty due to the aging population (Blatchford & Capewell, 1997; Sharkey & Gillam, 2010, Poteliakhoff & Thompson, 2011). Improvements in diagnosis and treatment of illness (Hobbs 1995); changes in incentives driven by the change in payment processes from a block grant to a tariff scheme (Farrar et al 2009, Information Centre 2010); a more open market, allowing private providers to treat NHS-funded patients (Naylor & Gregory 2009) and changes in working practices due to the new GP contract introduced in 2003 (Spurgeon 2005).

There is a lack of suitable data for studies of the relationship between primary and secondary care, so most analyses are based on small samples, and/or consider a few local areas. Cowling et al (2013), Cowling et al (2014), Cowling et al (2015) and Cecil et al (2016) found that better access to primary care is generally associated with less people seeking help at A&E departments. Gulliford (2002) shows that an increased supply of GPs is associated with lower hospital admissions and lower mortality across sixty-eight GP practices in North London; Harris (2011) finds that access to primary care does not explain differences in potentially avoidable emergency department attendances. Bankart et al (2011) and Gunther et al (2013)

identified some practice characteristics (shorter distance from hospital, a smaller 'list' size, an inability for patients to consult their preferred general practitioner) and patient characteristics that were associated with higher emergency admission rates, while Lowe et al (2005) study similar associations using Medicaid data. Wright and Ricketts (2010) show that in the US, having a higher proportion of primary care physicians amongst the local supply of physicians is associated with a lower number of emergency admissions at the metropolitan statistical area (MSA) level, but not when data is aggregated to the county level. O'Donnell (2000) concludes that the variation of referral rates among GPs remains largely unexplained. However, these studies do not consider emergency admissions, nor the link between emergency admissions and referrals. Furthermore, none of these papers explore the role of the GP gatekeeper and the effect on emergency admissions.

Our paper aims to provide a better understanding of various aspects of emergency admissions in England and its relationship to the provision of referrals. We utilise a dataset covering all English practices that treat NHS patients from 2004/5 to 2012/13 that allows comparisons between practices – both in cross-section and over time – of emergency admission rates and of GP referral rates. To the best of our knowledge, this is the only paper that use the total number of England GP practices to study this topic.

Using a panel data model with practice fixed effects, we estimate models of the referral rates and the hospital emergency admission rates of GP practices in England. The key result in our research is a positive relationship between referrals and emergency admissions, both in a simple cross-section comparison – practices with high referral rates also tend to have high emergency admission rates – and also in a model of practice level emergency admissions when referrals are positively correlated, after allowing for other influences on emergency admissions. We suggest that both of these observations reflect variation of patient health, across practices and within practices over time. We suggest that policy makers should aim to increase their

understanding of the interactions between these two types of healthcare in order to design appropriately focused policies.

Our results have important policy implications for healthcare, particularly as pressure grows on GPs to restrict referrals as a way of limiting NHS spending. If, as suggested, the simultaneous occurrence of higher referrals and above expected emergency admissions is due to variations over time in practice level ill health, simple constraints on referrals will reduce the flexibility of the system. Interventions to reduce referrals may reduce health system efficiency if patients are induced to increase reliance on emergency admissions. In particular, practices in deprived areas with a disproportionately large share of elderly patients with greater health needs tend to have more referrals and more emergency admissions, and as such may be more affected by policies to reduce referrals than other practices with different demographic features. Policy designers should consider this patient heterogeneity across practices and any practice targets to reduce referrals should vary accordingly. Although our results suggest referral rates are positively correlated with above-predicted emergency admission rates, this does not imply that a reduction in referral rates will reduce emergency admissions since we view both as driven by variations over time in health needs.

The rest of this paper is structured as follows. Section 2 describes the data and its sources. Section 3 describes the empirical strategy. Section 4 describes the results. Section 5 discusses potential explanations for these results and Section 6 concludes.

2. DATA AND DESCRIPTIVE STATISTICS

The data used in this study comes from three different sources combined to produce a comprehensive dataset containing information on practices, patients, and hospital admissions, for the period 2004/5 to 2012/13. The data required to perform this analysis in more recent

years is not available as the data now provided by NHS Digital is aggregated at practice level, without information about the residence of the patients or the type of contract possessed by individual GPs. We use data from the GP Workforce data, the Quality and Outcomes Framework (QOF) database, and Hospital Episode Statistics (HES). The unit of analysis is GP practices. GP data is primarily from the GP Workforce data, provided by the Health and Social Care Information Centre (HSCIC), and contains data on current General Medical Practitioners (GPs) including their gender, whether they studied in the UK or abroad, their years of qualification and birth, and the type of their employment contract. It also includes information about the practice where they work, plus start and end dates for other employment from 2004/5 until 2012/13 for each GP. This allows us to link practice data on the patient list size, the total number of GP hours at the practice, the total number of FTE GPs, the location, whether it is single handed, and the type of contract (GMS or PMS) agreed with the PCT. For each practice, we also link information about the characteristics of the registered patients provided by NHS Digital. This information includes the age structure (proportions in several age groups), proportion of men and women, and the number of registered patients resident in each LSOA (Lower Super Output Area).

Using information regarding the LSOA of patient residence for each practice, we are able to calculate the weighted average of the deprivation domain of the English Indices of Multiple Deprivation (IMD) at practice level. See Noble et al (2008) for a discussion of the IMD. The GP Practice IMD is estimated by taking a weighted average of the IMD scores for each LSOA in which the practice has patient registrations. The weights of the IMD scores are the percentage of the practice's registrations in each LSOA. Using an equivalent method, we also calculate an Index of Rurality for the patients of each practice, by taking a weighted average of the rural indicator provided by the Office of National Statistics (ONS) of the LSOA of residence for each patient registered with the practice.

Patient demand for admissions will be affected by their illnesses. Therefore, we use clinical information concerning the prevalence of specific diseases for the patients registered with the practice, plus data concerning quality of practice care, provided in the QOF database. By the end of the period, 90% of the population was registered with a GP included in QOF, and the observed illness prevalence rates are representative at the England level. There may be some under reporting bias, especially in the early years, but unfortunately, we do not have an alternative set of data to control for illness prevalence amongst the population for the time period under investigation.

To estimate models of GP referrals and emergency admissions rates we also use data from HES. This database provides case-level information for NHS hospitals on all admitted inpatients and outpatients, as well as A&E attendances. 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 HES record includes a set of information at patient level, concerning clinical information about diagnoses and operations; patient characteristics, such as age group, gender, ethnicity, area of residence and registered practice; further administrative information, including time waited, dates and methods of admission, and geographical discharge information. We focus on first referrals made by GPs and emergency admissions final episodes (inpatients) to avoid double counting of patients at different points in the treatment pathway. This dataset includes a variable detailing the patient's practice code, which enables HES data to be linked to that from other sources as described above.

Our main variable of interest is the rate of emergency admissions per 1000 patients registered at a GP practice. Patients can enter into emergency care at a hospital (a graphical representation is given in Figure 1 of Department of Health (2013) by several routes. The majority of emergency admissions in English hospitals come from patients who present at

Accident and Emergency (A&E) departments, either independently or via the 999-ambulance service, and are then admitted into hospital wards. Patients can also be referred after attending their GP or an outpatient hospital appointment, or on the advice of NHS Direct or NHS Walk-in centres. We include emergency admissions from all sources and do not distinguish between the sources in our analysis.

Table 1 reports descriptive statistics by year of the variables used in our estimated models. The average referral rate by practice increased by approximately 33% between 2004/5 (153.6 per 1000 patients) and 2012/13 (205.1 per 1000 patients), with most of the growth occurring between 2007/8 and 2009/10. This was much greater than the 14% increase in emergency admission rates over this time period. The variation across practices also increased, with the standard deviation of average referral rate increasing from 40.80 to 55.76. Emergency admissions were constant from 2004/5 until 2007/8 at 82-85 per 1,000 registered patients per GP practice, and from 2008/9 to 2012/13 when the average rate across practices was approximately 93, with a jump in-between the two groups of years. This trend is discussed by Imison and Naylor (2010) and coincides with a time of system reform, such as the new GP contract, which began in 2006, and the achievement of the 18-week elective waiting time target. GP referral behaviour has changed much more than hospital behaviour in admitting emergency patients and GP practice behaviour is becoming more diverse across England.

Patient demographics remained stable during this time but there were significant changes over time to the other practice characteristics we consider. The average list size grew by about 8% over the period 2004/5 to 2012/13, and the share of single-handed practices decreased from 20% in 2004 to 12% in 2012. Looking at GP characteristics, the average percentage of female GPs per practice increased from 30% to 39%, and the average percentage by practice of full time GPs fell from 82% to 67%; this may partly be a consequence of the increase in female GPs. A new GP contract allowed GPs to work less than 60% FTE from 2006 onwards and an

increasing proportion of doctors took up that option, reaching 15% by 2012. The average percentage of partner GPs has decreased from 92% to 83%, 2004–2012, while the mean age of GPs by practice increased by almost one year during this period.

TABLE 1 HERE

Table 2 shows how the descriptive statistics vary between the GP practices with the 30% most deprived and the 30% least deprived practices.

There are significantly more emergency admissions in deprived areas. However, GPs in both types of area have similar rates of referrals. This may well indicate too low referral rates in deprived areas. Less deprived practices have on average a larger proportion of older patients, with 24% aged 60 or above, compared to 16% in more deprived areas, and are more concentrated into rural areas.

GP and practice characteristics vary markedly in the two samples. Practices with patients from the more deprived areas tend to have smaller list sizes and fewer FTE GPs. They have on average older GPs, fewer female GPs, and a higher share who have been trained overseas. Additionally, there is a greater proportion of GPs who work full-time. There is also a significantly lower proportion of practices on the GMS contract in the "more deprived" sample.

Given these differences, it is possible that the relationship between referrals and emergency admissions could vary in areas with different levels of deprivation. We explore whether this is the case in Section 4.

TABLE 2 HERE

Figure 1 plots the cross–section relationship across practices between emergency admissions and GP referrals for 2004/5 and 2012/13. This is one of the key relationships considered in this paper. In 2004/5, GP practices have an annual mean of approximately 150 referrals and 75 emergency admissions per 1,000 patients. There appears to be a weakly

positive correlation, in which areas with higher referrals also tend to have slightly higher emergency admissions. The distribution in 2012/2013 has a similar shape, although the distribution moves upwards and to the right between 2004/5 and 2012/13.

FIGURE 1 HERE

In Figure 2 we present heat maps showing the relationship between referrals and emergencies in the 30% most deprived practices and the 30% least deprived practices respectively, in 2012/13. Deprivation is determined by a weighted average at each practice of the IMD of their registered patients, as described in the data section.

FIGURE 2 HERE

The tendency of practices that make more referrals to also have more emergency admissions is to be expected in simple comparisons of practices without controls: some practices – for what-ever reason – are likely to have less healthy patient populations, and this disparity of health between practices is a key determinant of both practice referrals and emergency admissions. It should be noted that this relationship between referrals and emergency admissions is disparate in deprived areas (left panel), where the behaviour of practices appears to be less similar to each other and more variation is present. In fact, there appear to be two mass points in the plot of "most deprived" practices, with one point at significantly higher emergency admissions and moderately higher referrals than the second mass point; the former may be thought of as "low engagement" with the health service. The more prosperous areas have a similar average referral rate to the less affluent areas but 50%-60% lower emergency admission rates. This suggests that practices at the two ends of the deprivation scale may react differently to any policy interventions.

3. EMPIRICAL STRATEGY

This work aims to improve understanding of how policy might influence the level of emergency admissions by studying two central and inter-related issues. Firstly, we study the determinants of the rate of practice first referrals; and secondly the determinants of the probability of emergency admissions at practice level given the referral rate, which can be affected by exogenous influences linked to policy as well as to the total rate of first referrals from the referring practice. By studying these two decisions, we are able to identify the effects of specific variables on the likelihood that referrals can influence emergency admissions.

Our empirical strategy consists of two steps: first, we estimate a model of how the characteristics of general practices, GPs, and their patients, influence the rate of referrals.

We estimate first the referral model for each practice in time *t*:

$$(1)\ lnR_{it} = \theta GP_{it} + bPat_{it} + cX_{it} + z_{it} + \sigma_i + \mu_t + \varepsilon_{it}$$

where lnR_{it} represents the logarithm of the number of referrals at each practice *i* in each year *t* per one thousand patients. Referrals made by others (specialists, nurses, etc.) are excluded in order that we capture the effects of various primary care characteristics on GP referral decisions, and the subsequent hospital admissions. GP_{it} is a vector of GP and practice characteristics that are time-varying at practice *i* in time *t*. The GP characteristics include the proportion of female GPs, mean practice GP age, ethnicity of GPs, type of GP contract (provider/partner versus salaried), and practice characteristics (single handed practice, QOF score, and GMS/PMS contract), and θ is a vector of the slope effects of these variables.

The term Pat_{it} represents a vector of patients' characteristics at practice *i* in year *t* (the share of patients in each of several age categories, gender share, rural practice and the deprivation index). The term z_{it} capture the proportions of patients registered at each practice with each of several chronic diseases indexed *i* (asthma, cancer, chronic obstructive pulmonary disease, coronary heart disease, diabetes, epilepsy, hypertension, hypothyroidism, left

ventricular failure, mental health and stroke). This information is taken from the QOF prevalence data. X_{it} is the bed occupancy at PCT level, included to capture the hospital capacity. Finally, we control for time (μ_t) with year effects, and practices (σ_i) with fixed effects. We use panel data fixed effects to estimate our model. The fixed effect component, σ_i, captures unobserved heterogeneity across practices that is fixed over time. Additionally, we correct the standard errors in order to control for heteroscedasticity. We cluster standard errors at practice level.

Finally, as our objective is to understand the relationship between emergency and referral rates at practice level, we estimate a regression model of emergency admissions, where we control for the referral rates at practice level. To specify this model, we begin by assuming that the probability of a referred patient from practice *i* being admitted in emergency is determined by the various characteristics associated with the practice (patient, practice, and GP), and the practice rate of referrals. In this case, if the proportion of referred patients from practice *i* in year *t* is R_{it}, then the emergency admission rates for each practice, Em_{it}, is estimated as follows:

$$(2)\ lnEm_{it} = \alpha lnR_{it} + Pat_{it} + \gamma GP_{it} + cX_{it} + z_{it} + \sigma_i + \mu_i + \varepsilon_{it}$$

lnR_{it} is the logarithm of referral rates of practice i in time t and where z_{it}, Pat_{it}, and GP_{it}, that are emergencies.

We consider four ways of interpreting the relationship between referrals and emergency admissions and seek to identify the more appropriate by sign value of α, the parameter estimating the influence of referrals on admissions. If we estimate that $\alpha < 0$, an increase in practice referrals is associated with a reduction in emergency admissions after allowing for the national trend in emergency admissions and mean practice admissions. This would be consistent with practice referrals acting as substitutes for emergency admissions, and policy intended to reduce hospital activity by suppressing referrals may encourage patients to switch from elective to emergency care.

If we estimate $\alpha = 0$, we can infer that referrals have no effect on emergency admissions and therefore policy to reduce referrals can be undertaken without adverse effects to patient health.

If the estimate is that $\alpha > 0$, we may assume that a simple substitution relationship is not the dominant one. Instead, both variables may be driven by variations over time in some other factor, and quite probably, patient health: a less healthy practice list in a given year leads to both more referrals and more emergency admissions. If variations in patient health are the predominant driver of the relationship between referrals and emergency care, the suppression of referrals *may* also result in some of the increase in unhealthy patients being offered more emergency admissions, but the relationship is less strong than a direct substitute relationship, as suggested by $\alpha < 0$. As well as variations in health, the introduction of system reforms over the time period studied, may influence the relationship between referrals and emergency admissions. System reform carried out in the period we study emphasised patient choice and health awareness in elective care, which could affect some groups of patients differently than others.

A problem with estimating the model in equation (2) is that the referrals made by GPs, rather than being exogenous, may be correlated with unobserved determinants of admissions, even after controlling for the time-constant differences between practices and socio-economic factors. For example, a contagious local illness will not be randomly assigned across geographic areas and such unobservable factors may increase both referrals and the error term in the emergency equation, leading to biased and inconsistent estimates. It could also be that that some of referrals result in hospital admissions and some of these admissions could require a readmission through A&E. In these situations, the use of instrumental variable estimation is preferred.

It is important that the instrument chosen to replace an endogenous variable is correlated with the endogenous regressor, but is only correlated with the dependent variable indirectly through its influence on the endogenous regressor.

The instrument we choose for referrals to control for the potential endogeneity is the first lag of the referrals in practice *i* in time *t*. The referrals made at *t-1* may not be correlated with the emergency admissions decisions on current year, but could influence the decision of GPs to refer this year, as usually referrals are partly based on previous experience and the historical health levels of patients. The time-lag of one year for referrals could be questionable, however when using lag2 and lag3 of referrals rates, the results are similar. Finally, we correct the standard errors in order to control for heteroscedasticity, and cluster at practice level.

4. RESULTS

In this section, we present results from linear regression (OLS) analysis and two stage least square (2SLS) analysis of practice referral rates and emergency admission rates.

4.1. Practice referrals

Table 3 presents three specifications of a model of referrals by GP practice, using various sets of control variables. In all regressions, we control for disease prevalence using QOF data, plus year and practice fixed effects. Additionally, Model 1 contains practice characteristics; Model 2 adds patient characteristics and Model 3 adds information relating to GPs. Most of these variables are common drivers of healthcare activity; others are intended to test specific ideas.

The natural log of the QOF score (ln QOF score) is included as a measure of practice "quality". It may be expected that practices of higher quality are capable of treating patients

with less need of secondary services in hospitals. Single-handed practices may use more referrals as they are by definition smaller and likely to have less facilities on site. The rurality measure and the proportion of beds occupied at local hospitals are included to capture access issues – patients in rural areas are likely to be further away from hospitals; if local hospitals are near capacity, there may be less chance of an admission given referral – both of these factors could influence the likelihood of a GP referring their patient. Older (compared to younger) populations and females (rather than males) tend to consume more healthcare and we also include disease prevalences to control for underlying population health which is likely to be a key driver of referrals. It has also been established that deprivation has a key role to play in determining healthcare usage. The characteristics of GPs could also play a role. We would expect older GPs to have a better understanding of their patients' benefits from referral due to their greater experience.

Most of the variables concerning practice characteristics have similar effects regardless of the other control variables included in the regression. The ln QOF score is negative but insignificant throughout, as is whether or not the practice holds a GMS contract, and the proportion of occupied beds at the local PCT hospitals. The IMD of practice patients is positively correlated with referrals, but is only significant (at the 10% level) in one of the three specifications. Single-handed practices refer significantly fewer patients to hospital, but this effect becomes positive (insignificant) when GP characteristics are considered in Model 3. The only practice characteristic that is significant throughout is that those in rural areas tend to refer at a lower level. The proportion of beds occupied in local hospitals, which might be thought to deter referrals, does not appear to be significant in these models.

Having a larger proportion of female patients leads a practice to have higher referral rates, and practices with more patients aged between 60-74 and 75-84 produce more referrals.

Increasing the proportion of patients aged 85+ reduces the referral rate, but this coefficient is not statistically significant.

In Model 3 we add several variables that examine average characteristics of practices' GPs. There is just one variable with a positive and statistically significant impact on referrals relative to the reference case – having a higher proportion of female GPs at the practice.

Practices with a higher proportion of older GPs tend to refer at lower levels. The coefficient is small, -0.001, but this result is significant at the 1% level. The proportion of foreign-born GPs also has a negative and significant impact on practice referrals. Practices with a larger share of clinicians who are partners, and who possibly have more interest in the reputation of the practice, tend to refer fewer patients, although this is only significant at the 10% level.

Model 3 is re-estimated for samples of the 30% least deprived practices and the 30% most deprived to explore whether these results are consistent across areas with different levels of deprivation, and presented as Models 4 and 5 in Table 3.

The results show that practices with more QOF points refer less in all three samples, but the coefficient is only statistically significant in the least deprived areas. A rural location does not appear to matter for practices in either the least or most deprived areas, despite a strong association in the complete sample. Practices with a GMS contract refer significantly less in deprived areas but not in prosperous areas.

The age distribution affects practice referrals differently in most and least deprived areas. The peak in least deprived areas is among patients aged 60-74, whereas in most deprived areas the slightly younger age group of 45-59 has the most referrals. Lifestyle and health issues probably cause this, with people from more deprived backgrounds generally experiencing worse health and less favourable habits.

Increasing the size of practice (by number of FTE GPs) reduces the referral rate in less deprived areas but has a positive (non-significant) effect in more deprived areas. As Table 2 shows, there tend to be greater numbers of foreign-born GPs in more deprived practices. Increasing the proportion of GPs with this characteristic at a practice leads to less referrals in the more deprived areas, but not less deprived areas. GPs who work between 60% and 100% FTE refer significantly more patients in more deprived areas but not in less deprived areas. Finally, a greater number of partner GPs is more important in the most deprived areas, with this variable having a negative and significant (at the 10% level) coefficient but an insignificant coefficient in less deprived areas.

The evidence presented here suggests that referrals can be influenced by many factors and as such, these factors should be taken into consideration when policy changes are considered. For example, if the number of GPs was to be increased, it appears important to ensure that they are incentivised to locate in more deprived areas to maximise the increase in patient benefit attainable from those extra practitioners.

TABLE 3 HERE

4.2. Emergency admissions

Having identified some of the major determinants of GP referrals, we now turn to emergency admission rates at GP practice level, and provide in Table 4 the results from the estimations of three models of emergency admission rates. The first column (Model 6) shows a specification using practice, patient and GP characteristics. In column 2 (Model 7) we add the rate of referrals at practice level. Model 8 uses a two stage least square model (2SLS) to take into account the endogeneity of referral rates, using the first lag of referrals in practice *i* in time *t* as an instrument for referrals. All of these regressions also contain controls for disease

prevalence using QOF data, plus year and practice fixed effects, although we only include statistically significant variables in the main tables due to space limitations.

The impact of referrals on emergency admissions is considered in Model 7, which adds as an additional explanatory variable the natural log of the rate of referrals per 1,000 patients at the practice. This variable is a positive and highly significant predictor of emergency admissions, with a coefficient of 0.252. The coefficients of most of the other control variables are not materially affected by the addition of referrals to the estimation.

The final column in Table 4 estimates a model of emergency admission rates using lagged referrals to control for endogeneity – to ensure that we are identifying the effect of referrals on emergency admissions and not the other way around. As such, there is one year less data available for the estimation, so the number of observations falls from 65,817 to 57,761. This does not significantly alter the results from Model 8 although the F statistic is greater than 10. The coefficient on referrals is similar, at 0.263 instead of 0.252.

TABLE 4 HERE

To help interpret the coefficients, Table 5 shows the predicted change in emergency admission rates for a one standard deviation change in some of the important control variables. Increasing the natural log of the referral rate by one standard deviation would have the largest impact amongst these variables, with a one standard increase resulting in an estimated increase in the annual admission rate of 5.89 per 1000 registered patients. Increasing the ln QoF score by one standard deviation would lead to a drop in admissions of 1.08 per 1000 patients. Reducing the index of deprivation by one standard deviation would reduce the admission rate by 1.45 per 1000 patients per annum, while changing the age structure of a practice by one standard deviation relative to the 0-14 age group would change admission rates by between 0.19 and -5.36 admissions per 1000 patients per annum. A one standard deviation reduction in the proportion of female patients would decrease emergency admission rates by 0.39 per 1000

patients per annum. Reducing the average age of a practice's GPs by one standard deviation would increase the emergency admission rate by 1.38 per 1000 patients per annum.

TALE 5 HERE

In Table 6 we reproduce the regressions of equations 1 and 2 using two sub-samples of GP practices, the 30% of practices with patients from the most deprived areas and the 30% of practices with patients from the least deprived areas, to see whether our results are affected by deprivation. This analysis helps to identify heterogeneity across practices.

A key result from Table 6 is that the relationship between referrals and emergency admissions is quite consistent across the three samples (least deprived, most deprived and all practices), with a coefficient between 0.20 and 0.264, that is significant at the 1% level in all specifications. We discuss this positive association further in Section 5.

Some of the control variables behave differently in the two – more deprived and less deprived – samples. In the most deprived areas, the relationship between the ln QOF score and emergency admission rates is clear cut – practices with lower QOF scores have significantly higher emergency admissions across all specifications. However, in least deprived areas the variable is only significant in the basic OLS specification, which does not contain referrals. Once average referrals are included, the QOF score becomes less important. This suggests a more complex relationship between QOF, referrals, emergency admissions and deprivation, with higher QOF scores resulting in less emergency admissions in more deprived areas, but less referrals in more affluent regions. More work is required to explain the important relationship between QOF and emergency admissions.

The proportion of GPs with a GMS contract has a positive (insignificant) role in the most deprived areas but is negative and significant, at least at the 10% level, in the two contemporaneous specifications using data from the least deprived practices.

The age distribution has stronger effects in the most deprived areas, with those between 15 and 74 having significantly less emergency admissions than the 0-15s, and those aged 85

and above having significantly more. The only age group that have significantly different admission rates to the under 15s in the less deprived areas are the 85+.

The proportion of female patients is only a statistically significant influence on emergency admissions at practices in the less deprived areas, although the coefficient is positive in all areas.

TABLE 6 HERE

5. DISCUSSION OF THE RELATIONSHIP BETWEEN EMERGENCY ADMISSIONS AND REFERRALS

A result that appears consistently in the analysis is that the annual level of emergency admissions of patients from a specific practice is positively correlated with annual practice referrals, even after controlling for practice and year fixed effects, features of GP practices, local population illness levels, and patient social characteristics.

An interpretation of the relationship between referrals and emergency care that corresponds with this correlation is that depending on their symptoms, circumstances and personal preferences, patients seeking treatment do so from either emergency or elective providers; some practice attendees are referred, and some attendees for emergency care are admitted to hospital. Following diagnostics and perhaps treatment, the care pathways end. If in a given year more (less) patients are unwell and seek care, then both referrals and emergency admissions will tend to increase (decrease), as the number of patients presenting in both environments will increase (decrease), and referrals and emergency admissions will be positively correlated, as we have found.

The positive coefficient on referrals in the emergency admissions equation is an indication that a straightforward substitution relationship between referrals and emergency admissions is not the dominant one in understanding this relationship at practice level.

An additional explanation for high referrals and high emergency admissions in an area is that local GPs are referring the wrong patients, so that those being referred do not need hospital treatment and patients who are sick are forced to access emergency services. In this setting, the most appropriate policy would be to reduce inappropriate referrals whilst encouraging GPs to increase referrals for patients before they seek emergency care.

A further possibility is that there is another time varying influence that causes practice referrals and above-expected emergency admissions to be positively correlated over time. System reform, such as patient choice, was designed to increase patient involvement in determining care, and to increase patient empowerment. This policy is likely to have increased patient demand across both primary and elective care, changing both patient and GP behaviour.

6. CONCLUSIONS

There is limited research into the relationship between referrals for planned care and hospital admissions for emergency care, and we contribute towards filling that gap in the literature. Our particular concern is to assess whether or not the evidence suggests that policy to achieve cost savings by restricting referrals is damaging to patient health. We present analysis of practice-level data, and produce a model of emergency hospital admissions in which the rate of referrals made by GPs is a key explanatory variable. In this model, we control for a number of variables, including practice and time fixed effects, which enables us to explore specifically the relationship between referrals and above expected emergency treatment. A

better understanding of this relationship could help the formulation of better policies that can increase the efficiency of resource use in health care provision.

The positive influence of referrals in the model of emergency admissions indicates that these two types of healthcare do not primarily act as straightforward substitutes, at least in the setting we study where we use aggregated referral and admission data at practice level. We suggest that this is because variations over time in patient health lead to practice level changes in both referrals and emergency admissions that are positively correlated.

One policy conclusion from this analysis is that attempts to cut costs by simply limiting referrals is likely flawed. Based on our analysis, limiting the flexibility in GPs' ability to refer will make the system less responsive to health needs, and quite probably, more expensive.

A further policy-relevant observation is that we find differences in the relationship between emergency admissions and referrals for areas at different levels of deprivation. More prosperous areas have average *referral* rates similar to the less affluent areas, but have emergency admission rates as much as 50%-60% lower. This suggests more affluent populations are healthier, and more able to gain the care they require through planned rather than emergency routes.

In order to more fully understand what is driving the positive correlation that we identify between emergency admissions and GP referrals, it is important to study further the interaction between treatment pathways in healthcare, and in this way disaggregate this study. Unfortunately, we do not have data that will allow us to perform this analysis.

Further work would be helpful if undertaken at more disaggregated levels, in particular looking at patients with specific health conditions, and including other relevant control variables, such as ethnicity and patients' distance from providers, which are also likely to influence GP referrals and emergency hospital admissions.

REFERENCES

Bankart, M. J. G., R. Baker, A. Rashid, M. Habiba, J. Banerjee, R. Hsu, S. Conroy, S. Agarwal, and A. Wilson. "Characteristics of general practices associated with emergency admission rates to hospital: a cross-sectional study." Emergency Medicine Journal 28, no. 7 (2011): 558-563.

Blatchford, Oliver, and Simon Capewell. "Emergency medical admissions: taking stock and planning for winter: We need more logic and more honesty." (1997): 1322-1323.

Cecil, Elizabeth, Alex Bottle, Thomas E. Cowling, Azeem Majeed, Ingrid Wolfe, and Sonia Saxena. "Primary care access, emergency department visits, and unplanned short hospitalizations in the UK." Pediatrics 137, no. 2 (2016): e20151492.

Cowling, T. E., Cecil, E. V., Soljak, M. A., Lee, J. T., Millett, C., Majeed, A., ... & Harris, M. J. (2013). Access to primary care and visits to emergency departments in England: a cross-sectional, population-based study. PloS one, 8(6), e66699.

Cowling, Thomas E., Matthew Harris, Hilary Watt, Michael Soljak, Emma Richards, Elinor Gunning, Alex Bottle, James Macinko, and Azeem Majeed. "Access to primary care and the route of emergency admission to hospital: retrospective analysis of national hospital administrative data." BMJ quality & safety (2015): bmjqs-2015.

Department of Health. "Emergency admissions to hospital: Managing the demand." (2013).

Farrar, Shelley, Deokhee Yi, Matt Sutton, Martin Chalkley, Jon Sussex, and Anthony Scott. "Has payment by results affected the way that English hospitals provide care? Difference-in-differences analysis." BMJ 339 (2009): b3047.

Gulliford, Martin C. "Availability of primary care doctors and population health in England: is there an association?." Journal of Public Health 24, no. 4 (2002): 252-254.

Gunther, Stephen, Nick Taub, Stephen Rogers, and Richard Baker. "What aspects of primary care predict emergency admission rates? A cross sectional study." BMC health services research 13, no. 1 (2013): 11.

Harris, Matthew J., Brijesh Patel, and Simon Bowen. "Primary care access and its relationship with emergency department utilisation: an observational, cross-sectional, ecological study." Br J Gen Pract 61, no. 593 (2011): e787-e793.

Hobbs, Richard. "Near patient testing in primary care." (1996): BMJ vol: 312, 263-264.

Imison, Candace, and Chris Naylor. "Referral management." Lessons for success. London: Kings Fund (2010).

Lowe, Robert A., A. Russell Localio, Donald F. Schwarz, Sankey Williams, Lucy Wolf Tuton, Staci Maroney, David Nicklin, Neil Goldfarb, Deneen D. Vojta, and Harold I. Feldman. "Association between primary care practice characteristics and emergency department use in a Medicaid managed care organization." Medical care 43, no. 8 (2005): 792-800.

Naylor, Chris, and Sarah Gregory. "Briefing. Independent sector treatment centres: The Kings Fund. 2009."

NHS Information Centre (2010) Healthcare resource groups 4 (HRG4). www.ic.nhs.uk/services/the-casemix-service/new-to-this-service/healthcare-resource-groups-4-hrg4. Accessed Sept 2018

Noble, M., McLennan, D., Wilkinson, K., Whitworth, A., Barnes, H. and Dibben, C. The English Indices of Deprivation 2007. London: Department for Communities and Local Government. (2008). Available from:

www.communities.gov.uk/publications/communities/indiciesdeprivation07

O'Donnell, Catherine A. "Variation in GP referral rates: what can we learn from the literature?" Family practice 17, no. 6 (2000): 462-471.

Poteliakhoff, E., and J. Thompson. "Emergency bed use: what the numbers tell us." London: The King's Fund (2011). http://www.kingsfund.org.uk/sites/files/kf/data-briefing-emergency-bed-use-what-the-numbers-tell-us-emmi-poteliakhoff-james-thompson-kings-fund-december-2011.pdf

Sharkey, Kerith, and Lynn Gillam. "Should patients with self-inflicted illness receive lower priority in access to healthcare resources? Mapping out the debate." Journal of Medical Ethics 36, no. 11 (2010): 661-665.

Spurgeon, Peter, Carolyn Hicks, Stephen Field, and Fred Barwell. "The new GMS contract: impact and implications for managing the changes." Health services management research 18, no. 2 (2005): 75-85.

Wright, David Bradley, and Thomas C. Ricketts III. "The road to efficiency? Re-examining the impact of the primary care physician workforce on health care utilization rates." Social science & medicine 70, no. 12 (2010): 2006-2010.