The Power of Data: Assessing Primary Care Performance Using Routinely Collected Emergency Department Data

Nicos Savva
Management Science & Operations, London Business School, nsavva@london.edu

Sandra Sülz
Erasmus School of Health Policy & Management, sulz@eshpm.eur.nl

Mark Kinirons
Guy's and St Thomas' NHS Foundation Trust and Kings College London,
mark.kinirons@gstt.nhs.uk, richard.leach@gstt.nhs.uk

Emer Sutherland
King's College Hospital,
emer.sutherland@nhs.net

Richard Leach
Guy's and St Thomas' NHS Foundation Trust and Kings College London,
mark.kinirons@gstt.nhs.uk, richard.leach@gstt.nhs.uk

This paper develops and validates a methodology to assess the impact of Primary Care Practices (PCPs) on Emergency Department (ED) demand using operational data from two geographically adjacent EDs in England. The analysis focuses on ED attendances for Ambulatory Care Sensitive (ACS) conditions, which are more appropriately and cost-effectively treated in non-ED settings. Employing a variance decomposition approach that controls for observed and unobserved heterogeneity, the study demonstrates that approximately 52% of the variance in ACS attendance proportions is attributable to systematic differences between PCPs, with these differences being operationally significant. The constructed PCP-level measure of systematic variation in ACS attendances correlates with patient feedback in NHS surveys, Quality and Outcomes Framework scores, and PCP staffing decisions. This measure outperforms prior metrics based on ACS admissions in identifying PCP variability and demonstrates robustness to alternative modeling choices. The methodology's modularity allows application to single or multiple hospital datasets, offering valuable insights for policymakers to evaluate PCP-level interventions and their downstream impacts on EDs.

Key words: Primary Care; Emergency Department; Variance decomposition
History: September 24, 2024

You can't manage what you can't measure.

Peter Drucker

1. Introduction

1.1. Context and Problem Statement

Emergency departments (EDs) across the developed world are reporting a significant increase in demand for their services (see Figure 1 for evidence from England and Berchet (2015) for evidence

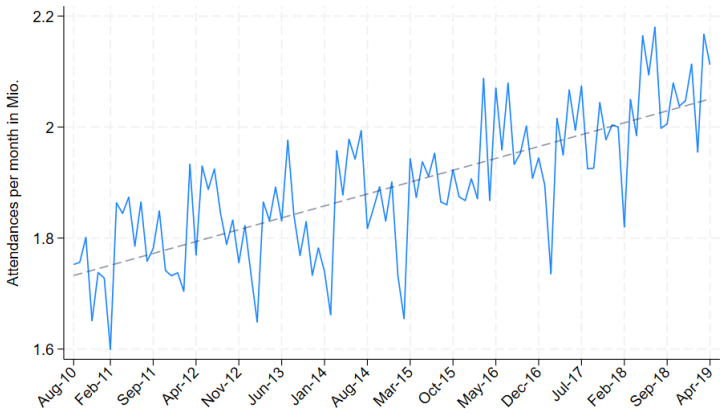

Figure 1 Emergency Department attendances in England, Aug 2010- Apr 2019. Source: Unify2 / SDCS data collections - WSitAE and MSitA

from around the world). While some of this increase can be attributed to demographic factors, such as the continuing rise in life expectancy leading to increasing numbers of older individuals with chronic conditions, estimates suggest that as many as 40% of ED attendances may be inappropriate and at least partially due to operational problems in primary care (Ismail et al. 2013).

Although many Primary Care Practices (PCPs)[1] provide an excellent range of services in a timely fashion, some are perceived as having unacceptably long waiting times for appointments, few options (if any) after regular office hours, or limited diagnostic resources that prompt patients to seek care at EDs (Berchet 2015, MacKichan et al. 2017, Bavafa et al. 2022). Treating such patients at the ED setting is considerably more expensive than providing care in non-emergency settings (Weinick et al. 2010, Galarraga et al. 2015), creates delays and worse outcomes for high-acuity patients (McCarthy et al. 2009, Soltani et al. 2022), and contributes to ED physician/nurse burnout (Watson et al. 2019).

From a system efficiency perspective, it is crucial to identify PCPs whose patients place a lower-than-average burden on EDs so that best practices can be identified and disseminated, and identify PCPs whose patients place a higher-than-average burden to address the underlying issues and provide support. However, assessing PCP performance in many healthcare systems is difficult due to lack of data and service fragmentation. PCP data is maintained locally using incompatible information systems, with complex and non-transparent data ownership structures, making direct comparisons between PCPs difficult (Gentil et al. 2017, Clarke et al. 2019).

In addition, PCP data do not typically include information on quoted waiting time for appointments and/or the number of patients that tried but failed to get an appointment. As a result, PCP performance is routinely assessed by means of patient surveys (Cowling et al. 2015). Surveys

[1] In this study, the abbreviation PCP refers Primary Care Practice (the business unit), not to individual physicians.

are by construction subjective, they can be slow and expensive to administer, and although they may provide valuable insights on patient perceptions, they cannot measure how variation in PCP performance affects ED demand.

1.2. Research Question and Methodology

This study aims to develop and validate a methodology to assess the impact of individual PCPs on ED demand using routinely collected operational data from two geographically adjacent EDs in England. The methodology quantifies the proportion of ED attendance variability attributable to PCPs, develops a measure of PCP performance based on ED utilization, and validates this measure against existing quality indicators. The analysis focuses on ED attendances for Ambulatory Care Sensitive (ACS) conditions, which are more appropriately and cost-effectively treated in non-ED settings. Employing a variance decomposition approach that controls for observed and unobserved heterogeneity, the study examines data over a period of five years from two geographically adjacent EDs belonging to two different hospital systems (NHS Trusts).

This study is set within the context of the National Health Service (NHS) in England, focusing on the relationship between Primary Care Practices (PCPs) and Emergency Departments (EDs). The NHS provides a unique setting for this research due to its universal coverage, free-at-point-of-service model, and the central role of PCPs in managing patient care. A comprehensive description of this setting, including the structure of primary and emergency care, patient decision-making processes, and the specific challenges faced by the NHS, is provided in Section 3.1.

1.3. Key Findings and Practical Implications

The study demonstrates that approximately 51.9% [95% CI: 37.8%, 65.7%] of the variance in ACS attendance proportions is attributable to systematic differences between PCPs, with these differences being operationally significant. The constructed PCP-level measure of systematic variation in ACS attendances correlates with patient feedback in NHS surveys, Quality and Outcomes Framework scores (a measure of how well PCPs comply with certain clinical and administrative guidelines used for financial reimbursement), and PCP staffing decisions. This measure outperforms prior metrics based on ACS admissions in identifying PCP variability and demonstrates robustness to alternative modeling choices.

The findings of this study have significant practical implications. If all underperforming PCPs were to improve to the 75th percentile, then approximately 6.7% of ACS attendances could be avoided, potentially reducing UK-wide ED attendances by 265,000 annually and saving the NHS £42 million per annum.

The methodology developed in this study offers valuable insights for policymakers to evaluate PCP-level interventions and their downstream impact on EDs. It can be applied in several ways:

1. Performance measurement for process improvement: By quantifying PCP performance, this approach enables health authorities to identify high and low performers, facilitating targeted investigations to uncover best practices and areas needing support.

2. Financial incentive design: A carefully implemented version of this methodology could be integrated into PCP reimbursement formulas, encouraging PCPs to internalize the costs associated with ED care for patients better served in primary care settings.

3. Resource allocation: Analysis suggests that using this methodology to inform resource allocation decisions, such as additional staffing, could lead to a greater reduction in ACS attendances at EDs compared to decisions based on other metrics.

The methodology's modularity allows application to single or multiple hospital datasets, offering flexibility in its use across various healthcare settings.

1.4. Contributions and Broader Implications

This study contributes to the growing body of empirical work in operations management and health services research, offering a novel approach to performance assessment that leverages routinely collected ED data.

A key strength of this methodology lies in its transparency, robustness to various modeling choices, and ease of implementation. While other statistical approaches may offer additional sophistication, this method strikes a balance between analytical rigor and practical applicability. It utilizes routinely collected data and employs interpretable techniques, making it accessible to healthcare managers and policymakers. This transparency and simplicity facilitate its adoption across different healthcare settings without requiring extensive additional data collection or complex analytical tools, enhancing its value for real-world application in healthcare management.

While developed within the NHS context, the methodology has broader implications. It can be adapted for various healthcare systems, including those with different funding models or organizational structures. For instance, in the United States, this approach could be valuable for integrated care providers like Kaiser Permanente, or for entities such as Medicare and Medicaid in assessing primary care performance. As healthcare systems worldwide face challenges like aging populations and resource constraints, this methodology could enhance care efficiency and improve patient experiences across diverse healthcare landscapes, from universal healthcare systems to insurance-based models.

2. Literature review

This paper builds on empirical work in operations management, particularly studies using observational data to measure performance in service systems. It also draws from and contributes to healthcare services research, focusing on primary care performance assessment and the use of

Ambulatory Care Sensitive (ACS) conditions as indicators. Finally, it extends the literature on provider profiling methodologies and the use of Emergency Department (ED) data in healthcare system analysis.

2.1. Empirical work in operations management

Recent years have seen a growing body of empirical work in operations management that leverages observational data to rigorously measure and assess performance in service systems. For instance, Tan and Netessine (2014) used restaurant software system data to establish that server workload affects meal duration and sales. In the banking sector, Staats and Gino (2012) and Xu et al. (2020) examined how repetition impacts worker productivity and how workload affects error rates, respectively. Ibanez and Toffel (2020) used data on restaurant inspection schedules to show that the order of inspections affects outcomes, while Wang and Zhou (2018) used supermarket checkout data to demonstrate that clerks in dedicated queues work faster than those in pooled queues, a finding echoed in the ED context by Song et al. (2015).

In healthcare settings, this approach has been particularly fruitful. Studies have established that ward- or provider-level workload affects patient service rates (KC and Terwiesch (2009), Berry Jaeker and Tucker (2017)), admission decisions KC and Terwiesch (2017), intensity of services provided Freeman et al. (2017), hospital reimbursement Powell et al. (2012), nurse absenteeism Green et al. (2013), and patient mortality rates Kuntz et al. (2015). Beyond workload effects, researchers have examined various aspects of healthcare operations. For example, Ibanez et al. (2018) used radiology data to study the impact of discretionary task ordering on worker productivity, while Kim et al. (2015) and Song et al. (2020) used admissions data to investigate how ICU admission delays and bed assignments affect patient length of stay. In the ED context, Song et al. (2018) showed how performance feedback influences worker productivity, and Batt and Terwiesch (2015) demonstrated the effects of queue length and waiting times on patient decisions to leave without being seen.As in this paper, the data exploited by these studies was routinely collected in the process of providing a service. In contrast to this paper, the work cited above uses the available data to study aspects of performance at the level of the unit in which the data was collected (e.g., ED data to assess aspects of ED performance).

More recent studies have adopted a system-wide perspective, examining how data collected at one level informs performance further downstream. For example, Freeman et al. (2021) investigated ED workload effects on hospital admission decisions, Soltani et al. (2022) explored ED workload impact on inpatient treatment efficiency, Song et al. (2022) analyzed how home healthcare visit duration affects hospital readmissions, and Bavafa et al. (2022) studied PCP availability effects on ED visits and readmissions. Similar to these studies, this work also takes a system perspective but

instead of using data collected at one unit to assess the performance of units further *downstream*, our work focuses on units that are further *upstream*. These are units that act as a precursor to the unit in which the data is collected.

2.2. Healthcare Services Research

Assessing primary care performance has been a focus of health economics and health services research for decades. The complexity of this task is evident in the diverse array of systematic reviews that have emerged, each examining different facets of the issue. These reviews have explored the impact of financial incentives Scott et al. (2011), Gibson et al. (2013), patient demographics Huntley et al. (2014), organizational structures and workforce dynamics van Loenen et al. (2014), Gibson et al. (2013), Huntley et al. (2014), and accessibility Rosano et al. (2013) on primary care performance.

Methodologically, researchers have employed a wide range of approaches, from qualitative interview studies Maisey et al. (2008), McDonald and Roland (2009) and patient surveys Cowling et al. (2016), Schoen et al. (2004), Campbell et al. (2009) to analyses of routinely collected quality indicators at the PCP level Brown and Lilford (2006), Downing et al. (2007), Hong et al. (2010), Dusheiko et al. (2011) and hospital admission data Harrison et al. (2014), Barker et al. (2017), Dusheiko et al. (2006), Vuik et al. (2017), Lavoie et al. (2019), Busby et al. (2017), Dusheiko et al. (2011). However, each of these methods has limitations. Qualitative interviews and surveys, while valuable for capturing patient perceptions, lack objective quantification. Quality indicators collected at the PCP level are susceptible to gaming behaviors Jürges and Köberlein (2015), Bastani et al. (2019). Hospital admission data can be influenced by hospital-specific variations in admission decisions Galarraga et al. (2015).

Our study proposes a novel approach: assessing primary care performance based on ED attendances. This method offers several advantages, including reflecting patients' revealed preferences rather than stated ones and being exogenous to both the hospital and the PCP, making it less susceptible to manipulation or gaming.

The use of ED attendance data for primary care performance evaluation is relatively unexplored, likely due to data availability constraints. However, two notable studies have paved the way. Dolton and Pathania (2016) used ED attendances to demonstrate the impact of extended PCP weekend services on emergency care demand. Dowd et al. (2014) developed a physician performance measure based on ED visits using Medicare claims data. Our study builds on these approaches but goes further in distinguishing between random variations within physicians and systematic differences between them, a critical distinction that Dowd et al. (2014) did not make.

2.3. Provider Profiling Methodologies

This work also contributes to the growing field of provider profiling, which aims to rigorously assess and visualize performance differences between healthcare providers (Jones and Spiegelhalter 2011, Spiegelhalter 2005, Racz and Sedransk 2010, Paddock 2014). This literature spans various healthcare contexts, including hospitals (Paddock et al. 2015), medical specialists (Adams et al. 2010), and PCPs (Thomas et al. 2004, Dowd et al. 2014). Recent studies have also explored novel data sources for performance assessment. For instance, Lu and Rui (2017) investigate the potential of using online physician reviews.

The literature on provider profiling highlights several challenges in selecting appropriate performance measures and applying rigorous risk adjustment (Weintraub and Garratt 2017, Baker and Chassin 2017, Braithwaite 2018). This study addresses some of these challenges through several methodological considerations. The measure of ACS attendance at the ED, while potentially subject to recording discretion, is hospital-specific rather than PCP-specific and would, on average, affect patients from all PCPs attending any given hospital equally. The use of hospital-time fixed effects controls for variation in coding practices across hospitals and over time. The proposed measure, being based on ED data, is not susceptible to gaming through upcoding or underreporting by PCPs, as they have no control over patient decisions beyond providing better service. Although covariates used for risk adjustment may be prone to recording errors, they are at least recorded consistently across all PCPs, mitigating discretionary measurement issues. Finally, the statistical methodology employed in this study explicitly decomposes variation into systematic and random components, addressing concerns about conflating stochastic fluctuations with genuine performance differences.

These methodological considerations align with calls in the literature for "a more systematic and transparent approach to risk adjustment methods and their rationale" (Braithwaite 2018) and for prudence in the application of such measures (Dowd et al. 2014). Consequently, this work is positioned primarily as a tool for performance improvement rather than punitive assessment.

3. Research Setting, Methodology and Results

3.1. Overview of the National Health Service (NHS) in England

Healthcare in England is provided by the National Health Service (NHS) and is funded through taxes, with residents accessing primary, emergency, and specialist care free of charge.

3.1.1. Primary Care in the NHS Primary care is delivered by General Practices (hereafter referred to as PCPs), which are financially independent entities contracted by the NHS. These PCPs offer a comprehensive range of services, including consultations, prevention, screening,

immunization, some diagnostic services, and minor treatments. They also serve as gatekeepers for diagnostic and other specialist services typically offered in hospital settings.

PCPs make autonomous investment and staffing decisions and are reimbursed through a risk-adjusted capitation system augmented by pay-for-performance elements (Roland and Guthrie 2016). To access care, UK residents need to be registered with a PCP. Until 2015, patients had to register with a PCP in whose catchment area they resided. This has been relaxed since then with patients being allowed to choose PCPs more freely (NHS 2016).

Patients typically access PCPs by appointment during their usual operating hours. These vary but for most practices they are between 08:00 and 18:30, Monday to Friday. Out-of-hours access to primary care varies and may include telephone service, out-of-hours clinics, or home visits (National Audit Office 2014).

3.1.2. Emergency Care in the NHS Emergency Care is provided in hospital EDs, such as the those examined in this study. EDs are typically open 24/7 and offer a wide range of services and are accessible without an appointment. Since 2005, the government has set a target for EDs to assess, diagnose, treat, and admit or discharge patients visiting EDs within four hours of arrival (Weber et al. 2012). This target was met for 95% of patients until 2015 but has since deteriorated.

3.1.3. Patient Decision-Making in the NHS The NHS allows individuals to seek care at the PCP or the hospital ED. This choice is influenced by various patient-specific and PCP-related factors (Huntley et al. 2014). MacKichan et al. (2017) identified several factors contributing to patients choosing EDs over PCPs, including complex PCP appointment systems, limited PCP appointment availability, lack of out-of-hours PCP care, and perceptions of superior care at EDs.

3.2. Ambulatory Sensitive Conditions

The aim of this study is to utilize routinely collected operational hospital data, and more specifically patient attendances at hospital EDs, to identify PCPs whose patients exert disproportionate pressure on ED resources, especially when these attendances are not medically warranted. For this reason, we focus on the subset of patients that present to the ED with ACS conditions. These are conditions where effective community care and case management can help to prevent (but not completely eliminate) the need for ED attendance or hospital admission (Busby et al. 2015, Blunt 2013). ACS conditions are categorized into acute, chronic, and vaccine-preventable conditions (see Appedix §XXX for a list of these conditions). Examples of acute conditions include ear, nose, and throat infections or urinary tract infections. Most often, these conditions can be treated effectively in an ambulatory-care setting provided the patient has timely access to a PCP appointment. Examples of chronic conditions include asthma, chronic obstructive pulmonary disease, and congestive heart failure. These conditions require monitoring in the community to ensure

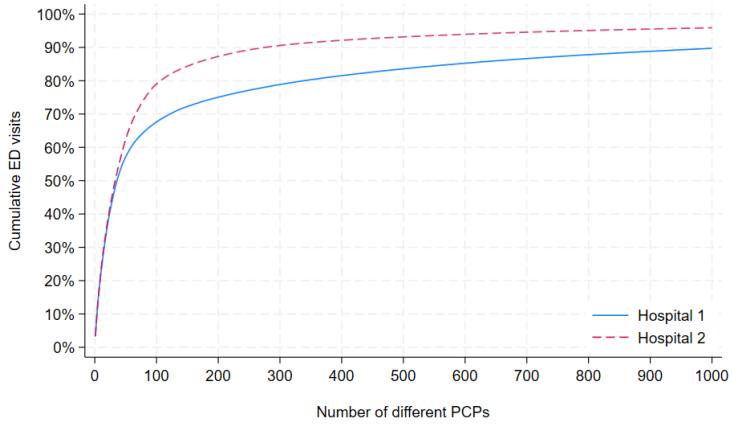

Figure 2 Proportion of ED attendances accounted for by different PCPs.

that the condition remains stable and the patient adheres to guidelines and medications. Examples of vaccine-preventable conditions include influenza and pneumonia. Their prevalence can be reduced or completely eliminated if an adequate vaccination scheme is in place. The premise is that patients registered with PCPs offering better care (e.g., timely access, effective case management) will exhibit fewer ED attendances for ACS conditions.

3.3. Data

The analysis uses approximately 1,400,000 patient ED attendances taking place at two large teaching hospitals in a major English metropolitan area over 5 years (2013–2017). The EDs are comparable in size, with about 380 daily patient attendances in 2017. Though geographically adjacent, they are part of different hospital systems (NHS trusts) and do not share any administrative resources.

The dataset includes arrival date/time, patient's registered PCP, presenting complaints, diagnosis, and patient disposition. This core data is supplemented with information from eleven publicly available sources that offer information on PCPs, NHS hospitals, and UK geolocation data (see Appendix §XX)

3.4. Sample Refinement: Focusing on Relevant PCPs

Since the goal is to estimate PCP performance, we have to exclude patients not registered with any PCPs (6.0% of observations) and those whose PCP couldn't be identified upon ED arrival (9.1% of observations). These exclusions likely remove overseas tourists or visitors who are not regular users of the UK healthcare system.

The remaining 84.9% of ED attendances are attributed to patients registered with 8,738 different PCPs across the country. However, as illustrated in Figure 2, ED demand is heavily concentrated among a small number of PCPs. Given the shape of this distribution, which is somwhat different for the two hospitals, the primary analysis focuses on the top PCPs accounting for 50% of attendances

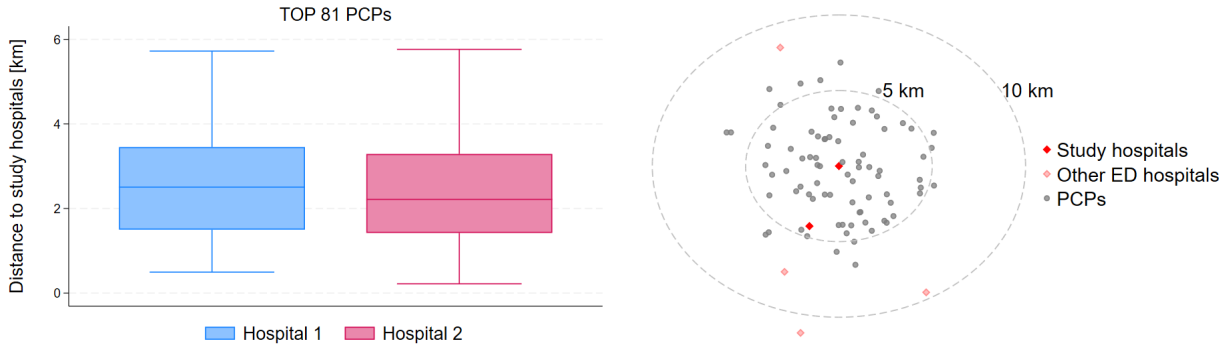

Figure 3 Interquartile range of distance and geographic location of top 81 PCPs.
To preserve anonymity of the ED and PCPs, the geographic coordinates have been transformed.

at Hospital 1 and 65% at Hospital 2. This approach identifies 81 distinct PCPs, with 8 PCPs appearing in the top lists for both hospitals due to their proximity (less than 5 km apart). As shown in Figure 3, all selected PCPs are within a 10 km radius of the study hospitals, with average distances of 2.6 km to Hospital 1 and 2.5 km to Hospital 2. To address potential concerns about the arbitrary nature of the cutoffs used (50% and 65%), sensitivity analyses with two different thresholds for each hospital are conducted in Appendix §3.1.

The exclusion of low-volume PCPs is crucial for the reliability of the analysis. For these PCPs at the distribution tails, there is insufficient data to reliably estimate performance, as their patients make only occasional use of the study hospitals. This approach aligns with other healthcare performance measurement strategies, such as the Hospital Readmission Reduction Program (HRRP), which only includes hospitals with at least 25 admissions in their readmissions measure (Chen and Savva 2018).

To account for potential operational disruptions, the study also excludes the final operating year for one PCP that closed during the study period, one PCP that started operating a new branch in the first year of the study period, data from the year a new branch started operating for another expanding PCP. As additional robustness checks, we also completely excluded PCPs with closures or expansions from the entire analysis (see Appendix §3.2). The final sample consists of 80 PCPs observed over 5 years.

3.5. Level of Aggregation and Unit of Analysis

Given that the data spans a period of five years and originates from two distinct hospital EDs, we need to make a decision as how to aggregate patient attendances in order to examine if there is systematic PCP-level heterogeneity.

Temporal Aggregation: From a temporal perspective, we want to be able to control for relevant time trends that affect all PCPs equally, but at the same time avoid constructing a

measure that is highly volatile and subject to seasonal variation. To address these concerns, patient attendances are aggregated at an annual level t. This annual aggregation allows for the control of time-related trends while smoothing out short-term fluctuations.

Organizational Aggregation: To account for potential systematic differences between hospitals (e.g., in coding practices), patient attendances are aggregated separately for each hospital ED (k). This approach helps control for hospital-level heterogeneity that could otherwise confound the analysis of PCP performance.

Specifically, for every PCP g we denote the count ACS attendances at hospital k in period t by:

$$P_{gkt} = \sum_{i=1, c \in ACS}^{N} X_{igkct},$$

where X_{igkct} is equal to 1 if patient $i \in \{1, ..., N\}$ is registered with practice $g \in \{1, ..., G\}$ in year t and presents to ED $k \in \{1, 2\}$ with a condition c that belongs to the set of ACS conditions and 0 otherwise. This aggregation yields an average annual count of 298 (SD = 143) ACS attendances per PCP at ED 1 and 227 (SD = 118) at ED 2.

Unit of Analysis and Cross-classified Panel Structure: Naturally, the number of ACS attendances P_{gkt} will exhibit variation that can be attributed to (i) stochastic fluctuations within PCPs over time, (ii) variation between PCPs that can be explained by observed heterogeneity (for example differences in the size of the practice, patient case mix, distance to the focal hospital and other hospitals in the area) or unobservable heterogeneity that has nothing to do with PCP service delivery (for example, differences in patient preferences over EDs of different hospitals, difference in hospital coding practices), and (iii) variation between practices that is due to persistent differences in PCP service delivery. Our primary focus is on the latter and in what follows we develop an econometric specification that aims to isolate this source of variability as far as possible.

To (at least partially) control for unobserved heterogeneity that is orthogonal to PCP performance we take advantage of the fact that in addition to patient attendances with ACS conditions we also have information on patient attendances with non-ACS conditions. These conditions, which include chest pain, suspected stroke, and trauma, require ED attendance and are less dependent on PCP service delivery compared to ACS conditions. If there exists heterogeneity (not related to service delivery) that makes patients from PCP i more likely to visit the ED of hospital k compared to patients from PCP j (e.g., because patients under PCP i may favor hospital k, whereas those under PCP j might opt for a different hospital), then, to the extent that these differences are common for patients with ACS and non-ACS conditions, both ACS and non-ACS attendances will be higher for PCP i compared to PCP j. Hence, we suggest gauging PCP performance by

normalizing ACS attendances by the overall number of patient attendances originating from PCP g in period t at ED k:

$$A_{gkt} = \frac{P_{gkt}}{\sum_{c \in C} \sum_{i=1}^{N} X_{igkct}}, \tag{1}$$

where X_{igkct} is defined as above and the set C includes all patient conditions (i.e., ACS and non-ACS conditions).

The proportion of ACS attendances A_{gkt} serves as the primary unit of analysis for this study. This creates a cross-classified panel data structure:

- PCP-level nesting: Observations are nested within PCPs (g). Each PCP has multiple observations over time and potentially across different hospitals.
- Hospital cross-classification: Some PCPs have data for only one hospital, while others have data for both hospitals. This means that the hospital dimension (k) is not strictly hierarchical but crosses with the PCP dimension.
- Time dimension: The annual-level (t) adds another dimension that applies across all PCPs and hospitals, but does not form a strict hierarchy with the hospital level.

This structure is cross-classified because PCPs can be associated with multiple hospitals, hospitals can be associated with multiple PCPs, and time applies across all PCPs and hospitals. Such cross-classified data structures are common in various fields. In educational research, for instance, students might be nested within both schools and neighborhoods, with no strict hierarchy between school and neighborhood levels (e.g. Goldstein 1994, Raudenbush 1993). Similarly, in medical research, patient observations over time at different healthcare organizations often form a comparable structure (e.g. Goldstein et al. 2002, Li et al. 2015). This cross-classified structure allows for the analysis of variability at multiple levels (PCP, hospital, and time) while accounting for the complex relationships between these levels.

3.6. Variance Decomposition

To identify the variation in the proportion of ACS attendances A_{gkt} that can be attributed to systematic differences between PCPs, we estimate the following model:

$$A_{gkt} = \alpha_0 + \alpha_C C_{gkt} + u_g + \epsilon_{gkt}. \tag{2}$$

This model accounts for any variation in the proportion of ACS attendances that is due to observed heterogeneity by controlling for a vector of (possibly time-varying) PCP- and hospital characteristics C_{gkt} on which we expand in §3.7.

The residual error is decomposed into two parts ($u_g + \epsilon_{gkt}$). u_g denotes the part of the variance that is explained by systematic differences between PCPs (also referred to as between-PCP variation). $\epsilon_{gkt} \sim N(0, \sigma^2)$ denotes the idiosyncratic error term (also referred to as within-PCP

variation). We make the usual strict exogeneity assumption, which can be expressed in terms of conditional expectations as $E(\epsilon_{gkt}|C_{gk1}, C_{gk2}, ..., C_{gkT}, u_g) = 0$ for all $t = 1, ..., T$, see Chapter 10 Wooldridge (2010)).

One possible choice in estimating the model of equation 2 and, in particular, the between-PCP variation u_g is to use a fixed-effect specification. Such specification has the advantage of allowing u_g to be arbitrarily correlated with the control variables C_{gkt}. However, since we only have 5 years of data and two hospitals (i.e., at most 10 data points per PCP) such an estimation would yield noisy estimates of PCP performance. Furthermore, the fixed-effect specification does not allow to control for any systematic differences across PCPs that remain unchanged over time (e.g., whether one of the study EDs is the closest to the PCP or not) as these would be collinear with the PCP fixed effect. Consequently, the estimated PCP fixed effect could become muddled with any time-invariant differences unrelated to the actual performance of the PCP.

To overcome both of these limitations we will use a random-effect model. More specifically, we make the additional parametric assumption that any heterogeneity between PCPs is drawn from Normal distribution ($u_g \sim N(0, \tau^2)$). This specification only requires one additional parameter to be estimated (τ^2) and allows to control for time-invariant heterogeneity between PCPs. However, these advantages come at a cost of requiring the additional identification assumption that the control variables C_{gkt} are orthogonal to the PCP random effect (i.e., $E(u_g|C_{gk1}, C_{gk2}, ..., C_{gkT}) = E(u_g) = 0$ see Chapter 10 Wooldridge (2010)). For example, this assumption may be violated if larger PCPs are more likely to make investments in resources that allow them to treat patients more effectively. In this case the PCP effect u_g would be correlated with the control for scale $Scale_{gt}$. Therefore, after we estimate the model we assess whether there is any evidence that this assumption is violated. In addition, as a robustness test, in Appendix §2.1 we also estimate the fixed-effect model which allows for arbitrary correlations between the control variables and the PCP effect.

Under these assumptions, the total variance of the random part of the model is equal to $\tau^2 + \sigma^2$ and the proportion of the variance that is explained by systematic differences between PCPs, often referred to as intra-class correlation (ICC), is given by $ICC = \tau^2/(\tau^2 + \sigma^2)$. A higher value of ICC suggests that more variability is due to systematic differences at the PCP-level, while a lower value suggests that most of the variability is due to random fluctuations within PCPs.

In estimating the model of (2) we report robust standard errors clustered at the PCP level. This modeling choice allows for errors to be heteroskedastic across PCP and correlated within PCP.

3.7. Controlling for Observable PCP and ED Heterogeneity

By defining the dependent variable as the proportion of ACS to total ED attendances (A_{gkt}) we control for the portion of PCP heterogeneity that affects ACS and non-ACS patients equally. In

this section, we further discuss controlling for observed and potentially time-varying heterogeneity, which may be important in cases where heterogeneity might affect ACS patients differently to non-ACS patients.

Scale and case-mix differences. We control for differences in scale and patient demographics between PCPs with the following variables: $Scale_{gt}$ which represents the number of registered patients (in 1,000), $Female_{gt}$ which indicates the proportion of female patients, and $Elderly_{gt}$ which denotes the proportion of patients aged 75 years and older. For most practices this information is reported twice a year, January and July, but for uniformity we use the January values.

In addition, we use the case mix index variable CMI_{gt} to control for additional differences in the patient case mix. Case-mix indices are used to measure the average severity of cases and to adjust reimbursement rates (Filistrucchi and Prüfer 2019). We calculate case-mix indices for the PCPs in our sample by relying on PCP reimbursement datasets that provide an adjustment to the number of patients registered by PCP by taking into account patients' needs.

Alternative hospital choice. One concern is that patients with ACS conditions from different PCPs may find visiting the study hospitals more (less) attractive depending on whether they are located close to (far from) the hospital and/or whether this is the closest suitable hospital to them. Normalizing the number of ACS attendances by dividing with the number of non-ACS attendances potentially addresses this concern but we further control for proximity of PCP g to study hospital k based on the distance between the two. More specifically, for each PCP g we determine whether the study hospital k is the closest hospital ($closest_{gk}$ equal to 1) or whether any alternative hospital is closer ($closest_{gk}$ equal to 0). All distance measures are calculated as straight line distances between the hospital ED and the PCP postcodes using the Stata command *geonear* (Picard 2012). Note that we measure proximity based on the location of the PCP and not the patient's home because the exact patient address is not available. Nevertheless, most patients register with a PCP close to where they live. We believe that using this binary measure instead of the actual distance between PCPs and the study hospitals is a more sensible modeling choice as this takes into account that some patients may be willing to travel a longer distance to get to the hospital simply because there is no closer alternative.

Socio-economic factors. Prior literature has shown that ACS attendance and admission rates are affected by patients' socioeconomic status with disproportionately higher ED attendance rates reported for lower socio-economic status (Oster and Bindman 2003, Johnson et al. 2012). Therefore, we control for socioeconomic differences between PCP locations. We do so via the index of multiple deprivation provided by the Department for Communities and Local Government. This index D_g captures dimensions such as average income, employment rates, education level, health characteristics, crime, barriers to housing and services. It is highly localized (at the postcode level)

and ranks locations in England from 1 (most deprived area) to 32,844 (least deprived area). We re-scale the index to capture ranks in units of 1,000. While deprivation is in theory a time-varying variable, it varies slowly and the data is only updated approximately every 5 years, which makes it a time-invariant factor in our analysis.

Hospital differences over time. We control for structural differences between EDs (e.g., differences in coding practices) and temporal variation within EDs through interacting year dummy variables $Year_t$ with ED_i dummy variables. It is essential to include these time variables. As noted by Wooldridge (2010), omitting them can induce serial correlation in the error term.

Importantly, in this methodology we do not control for operational or clinical differences between PCPs that are under the direct control of the PCP (e.g., opening hours, patient-to-staff ratio, expertise of clinical staff, diagnostic facilities, etc). We do not include these factors as they are the driving force behind the variability in PCP performance that we would like to capture. Indeed, we would like to have a measure of PCP performance that allows us to identify how such operational and clinical factors affect performance.

Full descriptive statistics by year of the variables used are provided in Appendix $ XX One observation is that there is relatively little variation between years compared to variation within a year, suggesting that most of these variables vary slowly over time.

3.8. Quantifying PCP systematic variability

We estimate the model described above using the *mixed* command, Stata Version 18. Table 1 Columns (1)-(3) provides the model estimates for different specifications relating to equation (2). Column (1) presents the results of a model without any controls except for the ED-year fixed effects. Column (2) adds controls for PCP location and deprivation. Column (3) adds controls for PCP scale, proportion of female and elderly patients, and case mix (CMI). All models suggest that more than half of the variance in PCP performance is systematic at the PCP level, with the ICC varying from 65.5% in the model without controls (Column 1) to 51.9% in model where we control for observable PCP differences (Column 3).

Figure 4 presents graphically the between-PCP variation ($\widehat{u}_g$) and the within variation ($\widehat{e}_{gt}$) based on the model of Column (3). Using this decomposition, one can think of PCPs with $\widehat{u}_g \leq 0$ as placing less burden on local hospital EDs than average (once we control for observable differences across PCPs and EDs) and, conversely, PCPs with $\widehat{u}_g > 0$ as placing more than average burden on EDs. The magnitude of the between-PCP variation in ACS attendances is not only statistically significant but also of practical importance. All things being equal, the difference in the proportion of ACS attendances between a PCP that is one standard deviation above the mean (i.e., a PCP with $\widehat{u}_g = 0.013$) and a PCP one standard deviation below the mean (i.e., a PCP with $\widehat{u}_g =$

Table 1 Quantifying PCP performance with variance decomposition

	ACS attendances			ACS admissions
	(1)	(2)	(3)	(4)
	A_{gkt}	A_{gkt}	A_{gkt}	Adm_{gkt}
Closest hospital		-0.001	-0.002	-0.009*
		(0.003)	(0.003)	(0.005)
Deprivation rank		-0.001***	-0.001***	-0.001
		(0.000)	(0.000)	(0.000)
Scale			0.000	-0.000
			(0.000)	(0.000)
Female			0.137	0.119**
			(0.101)	(0.047)
Elderly			-0.170	0.262
			(0.161)	(0.190)
CMI			0.061*	0.068*
			(0.033)	(0.040)
Constant	Yes	Yes	Yes	Yes
ED × Year FE	Yes	Yes	Yes	Yes
$\widehat{\tau}^2$	0.0003007	0.0002505	0.0001722	0.0001519
$\widehat{\tau}^2$ 95 CI	[.0001418; .0006376]	[.0001125; .0005578]	[.0001038; .0002856]	[.0000862; .0002675]
ICC	65.5%	61.3%	51.9%	17.3%
ICC 95 CI	[46.9% ; 80.4%]	[41.1% ; 78.3%]	[37.8% ; 65.7%]	[9.9% ; 28.4%]
Model Wald χ^2	570.9	862.2	916.5	379.2
Prop $> \chi^2$	0.000	0.000	0.000	0.000
Log Pseudolikelihood	1193.0	1199.6	1182.1	905.4
Observations	436	436	426	426
Number of groups	80	80	80	80

PCP-clustered standard errors in parentheses.
*** p<0.01 ** p<0.05, * p<0.1.
Note. Number of observations differ due to missing case-mix information in 10 PCP-years in Column (3) and in Column (4).

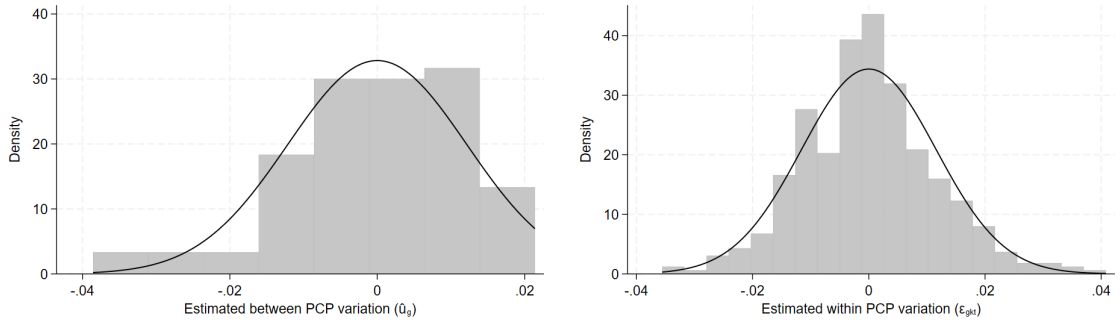

Figure 4 Variation in PCP performance: between PCPs (left), within PCPs (right)

−0.013) is 2.6% -points, which is equivalent to a 15.4% difference relative to the proportion of ACS attendances in the sample (0.169). Since the average PCP has 253 ACS attendances per annum per ED, this translates in a difference of 39 ACS attendances per annum per ED by each PCP.

We conclude by evaluating the random-effects model assumptions: i) Normal distribution of unobserved between-PCPs variation u_g, and ii) orthogonality of u_g to control variables.

A skewness-kurtosis test (p=0.020) suggests some deviation from normality for $(\widehat{u}_g)$. However, the predicted $(\widehat{u}_g)$ distribution need not match the true u_g distribution (McCulloch and Neuhaus

2011b). Even with misspecification, random effect variance estimates generally remain unbiased (McCulloch and Neuhaus 2011a, Bell et al. 2019, Maas and Hox 2004), and PCP ranking consistency is maintained for similar cluster sizes (McCulloch and Neuhaus 2011b).

To address potential violations of the orthogonality assumption, we employ Raudenbush's adaptive centering approach (Raudenbush 2009). This method involves transforming time-varying covariates by centering them around cluster means across multiple dimensions (in our case, PCP and hospital). This transformation helps to separate within-cluster effects from between-cluster effects, potentially mitigating bias from time-invariant confounding. We estimate an alternative model using these transformed variables for time-varying PCP covariates (see Appendix §xx). Comparing ($\widehat{u}_g$) from this alternative specification with our primary model shows no significant distributional differences (Kolmogorov-Smirnov test $p = 1.000$) and high correlation (0.920, p< 0.001). This suggests ($\widehat{u}_g$) is robust to time-invariant confounding from observable PCP covariances. However, we acknowledge that this approach does not directly test whether the true u_g is uncorrelated with the observable PCP covariates and cannot eliminate the possibility of confounding from unobservable PCP characteristics. To further address these concerns, we estimate a fixed-effects model (Appendix §2.1), which allows for arbitrary correlation between PCP effects and observable characteristics. This model yields similar results, with 77.5% of variation attributed to systematic PCP differences and high correlation (0.751, p$<$0.001) between fixed and random effect PCP-effect estimates.

Lastly, we find no evidence of non-linear scale effects on ACS attendance proportions (Appendix §2.2).

4. Validation of the PCP measure

By deploying the methodology outlined in the previous section we are able to construct a measure $\widehat{u}_g$ that captures the systematic variability between PCPs in the burden their patients place on local EDs. In this section we seek to establish whether this measure $\widehat{u}_g$ is indeed related to PCP's service delivery as opposed to other factors (e.g., residual unobserved heterogeneity). We do so by examining whether it is correlated with i) what patients say about their PCPs in the annual NHS survey; ii) the score PCPs receive in the Quality of Outcomes Framework; iii) staffing decisions made by the PCPs. In all cases we find positive evidence.

4.1. Correlation with patient survey outcomes

Every year, the NHS conducts a survey of over two million people about their PCP experiences (NHS 2023).The surveys consist of a number of questions relating to access to PCP services, the patient's experience with their PCP appointments, management of care plans, and a general assessment of the overall experience with their PCP. We focus on two surveyed items: (i) whether

patients would not recommend their PCP, and (ii) whether patients experienced access problems. These items were chosen as they are consistently surveyed across the eyears and capture overall dissatisfaction and access issues, respectively.

The patient survey data is reported biannually (July/June and December) until 2016, and annually from 2016 onwards (July/June). Therefore we use the July/June survey. For each PCP g in year t, we define:

- P^1_{gt}: proportion of patients who would not recommend their PCP (mean=8.7%, SD=0.067)
- P^2_{gt}: proportion of patients unable to make an appointment with their PCP (mean=13.4%, SD=0.062)

These measures are positively correlated (0.629, $p < 0.001$), indicating similar but not identical patient perceptions.

To examine the relationship between PCP variability in ED attendances ($\widehat{u}g$) and these survey measures (P^igt), we estimate:

$$P^i_{gt} = \beta_0 + \beta_U \widehat{u}_g + \beta_\epsilon \widehat{\epsilon}_{gt} + \beta_C C_{gt} + u^P_g + \epsilon^P_{gt}. \tag{3}$$

The measures $\widehat{u}_g$ and $\widehat{\epsilon}_{gt}$ are estimated using model (2), with $\widehat{\epsilon}_{gt}$ averaged for PCPs with multiple observations per year. We note that since $\widehat{u}_g$ and $\widehat{\epsilon}_{gt}$ are model estimated measures, they are measured with error. If classical errors-in-variables assumptions hold, the model's estimated coefficients (an in particular β_U) will be conservative. The vector C_{gt} incudes the same PCP-level controls as that of model (2). We decompose the error into between-PCP (u^P_g) and within-PCP (ϵ^P_{gt}) components, clustering at the PCP level. If the estimated measure $\widehat{u}_g$ relates to PCP service delivery we would expect the coefficient $\beta_U > 0$.

Table 2, Column (1)-(2), presents the results. We find that between-PCP variation ($\widehat{u}_g$) is positively associated with both survey measures. Relative to the sample mean, a one standard deviation increase in a PCP's proportion of ACS attendances corresponds to:

- 18.6% (95% CI [4..9%, 32.4%]) increase in patients not recommending their PCP
- 10.0% (95% CI [3.0%, 16.9%]) increase in patients reporting access problems

In contrast, within-PCP variation ($\widehat{\epsilon}_{gt}$) shows a more nuanced pattern: it has zero effect on patients' likelihood to recommend their PCP, but a positive, albeit smaller, effect on reported access problems. This differential impact suggests that while $\widehat{\epsilon}gt$ may capture some short-term fluctuations in access-related performance, it does not consistently reflect broader quality perceptions, unlike the more stable between-PCP variation ($\widehat{u}_g$).

These results provide strong evidence that $\widehat{u}_g$ is significantly related to patient-perceived PCP quality. The $\widehat{u}_g$ measure offers several advantages over traditional survey-based approaches. It is

Table 2 Validation: Correlation with patient survey outcomes

	Performance measure based on ACS attendances		Performance measure based on Average ACS rates		Performance measure based on ACS admissions	
	(1) Not Recommending	(2) Had No Access	(3) Not Recommending	(4) Had No Access	(5) Not Recommending	(6) Had No Access
$\widehat{u}_g$	1.235***	1.039***	0.051	0.143	0.295	0.632
	(0.464)	(0.368)	(0.189)	(0.148)	(0.594)	(0.543)
$\widehat{\epsilon}_{gt}$	-0.008	0.516**			0.056	0.104
	(0.216)	(0.240)			(0.087)	(0.105)
Closest hospital	-0.008	-0.006	-0.007	-0.005	-0.007	-0.005
	(0.016)	(0.014)	(0.015)	(0.014)	(0.015)	(0.014)
Deprivation rank	-0.001	-0.004***	-0.001	-0.003***	-0.001	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Scale	-0.002	-0.000	-0.002*	-0.000	-0.002*	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Female	-0.141	0.305***	-0.118	0.287**	-0.115	0.287**
	(0.123)	(0.116)	(0.128)	(0.116)	(0.134)	(0.127)
Elderly	1.670***	1.349***	1.508***	1.240***	1.503***	1.234***
	(0.496)	(0.445)	(0.491)	(0.444)	(0.487)	(0.433)
CMI	-0.246**	-0.304***	-0.236**	-0.297***	-0.237**	-0.298***
	(0.106)	(0.087)	(0.109)	(0.089)	(0.108)	(0.090)
Constant	Yes	Yes	Yes	Yes	Yes	Yes
Year × ED FE	Yes	Yes	Year FE	Year FE	Yes	Yes
PCP effect	Random	Random	Random	Random	Random	Random
Wald χ^2	21.4	63.9	21.3	35.1	20.0	37.3
Prop χ^2	0.045	0.000	0.031	0.000	0.067	0.000
Log. Pseudolikelihood	614.5	607.6	611.7	601.1	612.0	602.1
Observations	387	387	387	387	387	387
Number of groups	80	80	80	80	80	80

PCP-clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
Note. The columns "Not Recommending" denote the proportion of patients that would not recommend their PCP to others. The columns "Had No Access" denote the proportion of patients reporting problems with accessing their PCP. $\widehat{u}_g, \widehat{\epsilon}_{gt}$ based on ACS attendances in Column (1)-(2), average ACS rates in Column (3)-(4) and ACS admissions in Column (5)-(6).

derived from actual ED attendance data, making it more objective and less susceptible to subjective patient biases or recall errors that often influence surveys. Unlike surveys, which may be affected by temporal factors or individual mood, $\widehat{u}_g$ provides a more consistent measure of PCP performance over time. Moreover, while surveys capture patient perceptions, $\widehat{u}_g$ reflects the actual impact of PCP performance on the healthcare system, specifically ED utilization. Crucially, $\widehat{u}_g$ allows for precise quantification of individual PCP burden on local EDs, a feature not possible with survey data alone. This enables sophisticated cost-saving analyses, such as estimating potential ED savings from improving underperforming PCPs (see §6), providing actionable insights for healthcare system optimization. Additionally, once established, this methodology can be applied system-wide with minimal additional data collection, unlike resource-intensive patient surveys. In sum, while patient surveys remain valuable for understanding subjective experiences, the $\widehat{u}_g$ measure provides a more robust, system-level indicator of PCP performance with direct implications for healthcare resource allocation and policy-making.

4.2. Correlation with Quality of Outcomes Framework scores

To further validate whether the between-PCP variability in unwarranted ED visits ($\widehat{u}_g$) reflects PCP service quality, we examine its correlation with the Quality and Outcomes Framework (QOF) score. The QOF is a voluntary incentive scheme rewarding PCPs for quality care, with high participation rates (94.8% of English PCPs in 2017, including all PCPs in this study) (NHS 2018). It covers various clinical and public health domains, with indicators adjusted annually. Higher QOF scores have been associated with improved clinical outcomes, including reduced mortality rates (Ahmed et al. 2021).

Given that QOF scores are generally high (average 537.5 out of 559 points in 2017) (NHS 2018) and vary yearly, we focus on the proportion of points achieved (QOF_{gt}) rather than absolute scores. In our sample, the median QOF_{gt} increased from 95.9% in 2013 to 97.2% in 2017. If $\widehat{u}_g$ indeed reflects PCP care quality, we expect a negative correlation with $QOFgt$.

Due to the skewed distribution of QOF_{gt} (Figure 5a), we employ a binary choice model, categorizing PCPs into those achieving ≥90% or ≥95% of QOF points. We estimate a binary choice (mixed-effect Probit) model using the same controls and error structure as the model of equation (3):

$$\begin{aligned} HighQ^*_{gt} &= \gamma_0 + \gamma_U \widehat{u}_g + \gamma_\epsilon \widehat{\epsilon}_{gt} + \gamma_C C_{gt} + u^Q_g + \epsilon^Q_{gt}, \\ HighQ_{gt} &= 1[HighQ^*_{gt} > 0]. \end{aligned} \tag{4}$$

Results in Table 3 show a negative coefficient γ_U for both thresholds, though it's statistically significant only for the 90% threshold. PCPs performing one standard deviation below the mean have a 92.3% (95% CI [88.5%; 96.8%]) probability of achieving ≥90% QOF points, compared to 81.3% (95% CI [72.3%; 89.6%]) for those one standard deviation above the mean. However, there is considerable estimation uncertainty in these figures, as evidenced by the large standard errors. This uncertainty suggests that while there is a clear trend, the relationship between $\widehat{u}_g$ and QOF scores may be more nuanced than a simple linear association. Specifically, the data indicates that $\widehat{u}_g$ is a stronger predictor of very poor QOF performance (i.e., less than 90% QOF score) but its predictive power diminishes when distinguishing between PCPs in the higher ranges of QOF scores (e.g., above 90% or 95% of QOF score). This suggests that $\widehat{u}_g$ is particularly effective at identifying the worst-performing PCPs in terms of care quality, as measured by QOF scores.

We also observe an unexpected positive relationship between the idiosyncratic error term $\widehat{\epsilon}_{gt}$ and achieving very high QOF scores (≥95%). This result warrants further investigation with larger datasets or alternative model specifications to determine if it represents a meaningful pattern or is an artifact of the current data structure or model.

Figure 5b further demonstrates this relationship through cumulative distribution functions (CDFs) of $\widehat{u}_g$ for PCPs achieving ≥90% QOF points versus those that do not. The near-perfect stochastic dominance supports that $\widehat{u}_g$ reflects PCP service quality.

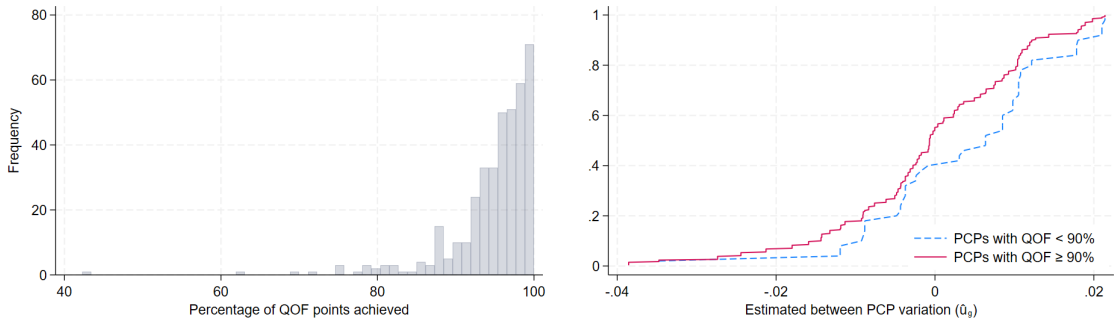

Figure 5 Distribution of QOF_{gt} and relationship with $\widehat{u}_g$

Table 3 Validation: Correlation with QOF scores

	Performance measure based on ACS attendances		Performance measure based on Average ACS rates		Performance measure based on ACS admissions	
	(1)	(2)	(3)	(4)	(5)	(6)
	$P(QOF_{gt} \geq 90)$	$P(QOF_{gt} \geq 95)$	$P(QOF_{gt} \geq 90)$	$P(QOF_{gt} \geq 95)$	$P(QOF_{gt} \geq 90)$	$P(QOF_{gt} \geq 95)$
$\widehat{u}_g$	-38.993**	-10.278	1.900	2.965	-30.095	0.293
	(15.638)	(17.269)	(6.144)	(6.063)	(21.478)	(23.167)
$\widehat{\epsilon}_{gt}$	12.846	19.156**			-1.203	-0.830
	(11.501)	(8.774)			(5.700)	(3.897)
Closest hospital	0.904	0.790	0.901	0.769	0.904	0.778
	(0.690)	(0.549)	(0.697)	(0.535)	(0.703)	(0.536)
Deprivation rank	-0.028	-0.018	-0.023	-0.012	-0.029	-0.018
	(0.038)	(0.042)	(0.042)	(0.043)	(0.039)	(0.041)
Scale	-0.021	-0.040	-0.003	-0.033	-0.004	-0.033
	(0.055)	(0.065)	(0.057)	(0.066)	(0.056)	(0.065)
Female	6.531*	7.614	3.667	6.712	3.953	6.558
	(3.800)	(5.665)	(2.933)	(5.423)	(3.033)	(5.568)
Elderly	-24.666	7.115	-16.461	9.879	-17.687	9.953
	(19.268)	(18.168)	(18.700)	(17.936)	(19.353)	(17.953)
CMI	-1.036	-4.938	-1.475	-4.894	-1.360	-4.977
	(3.857)	(3.837)	(3.843)	(3.829)	(3.861)	(3.825)
Constant	Yes	Yes	Yes	Yes	Yes	Yes
Year times ED FE	Yes	Yes	Year FE	Year FE	Yes	Yes
PCP effect	Random	Random	Random	Random	Random	Random
Wald χ^2	26.4	25.5	18.9	21.5	23.3	23.5
Prop χ^2	0.010	0.013	0.062	0.028	0.026	0.024
Log. Pseudolikelihood	-114.7	-201.8	-117.3	-204.4	-116.5	-204.5
Observations	389	389	389	389	389	389
Number of groups	80	80	80	80	80	80

PCP-clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
Note.$\widehat{u}_g, \widehat{\epsilon}_{gt}$ based on ACS attendances in Column (1)-(2), average ACS rates in Column (3)-(4) and ACS admissions in Column (5)-(6).

A key advantage of the estimated $\widehat{u}_g$ measure is its resilience to gaming, a concern sometimes raised with the data currently used to estimate the QOF score (Ashworth and Kordowicz 2010). Being based on ED observations rather than PCP-reported data, $\widehat{u}_g$ is less susceptible to manipulation. This feature makes it a potential candidate for inclusion in QOF calculations, allowing for financial rewards to PCPs whose patients place lower burdens on local EDs. Moreover, $\widehat{u}g$ provides a system-wide perspective on PCP performance by capturing the downstream effects of primary

care on emergency services. Unlike the QOF score's focus on specific clinical indicators, $\widehat{u}g$ reflects a PCP's overall effectiveness in preventing unnecessary ED visits, complementing the QOF by addressing broader aspects of care quality. Importantly, as $\widehat{u}_g$ is derived from routinely collected ED data, it could provide a cost-effective way to augment existing quality measures without imposing additional reporting burden on PCPs.

4.3. Correlation with PCP staffing

To further validate that the estimated $\widehat{u}_g$ measure reflects PCP service delivery, we examine its relationship with the ratio of registered patients to full-time-equivalent (FTE) physicians. Higher patient-to-staff ratios have been linked to lower quality care in hospitals (Kane et al. 2007, Needleman et al. 2002, Pronovost et al. 2002) and primary care settings (Chang et al. 2011). We hypothesize that PCPs with higher ratios may have longer wait times and potentially lower care quality, leading to increased ED burden.

We measure patient-to-staff ratio by calculating the average number of patients per FTE physician employee in practice g throughout the study period. One PCP exhibited implausibly low staffing data for at least one year (0.329 FTE physician for 3,565 patients), which we attribute to a likely data entry error. We exclude this PCP from our analysis. The remaining PCPs show considerable variance in their patient-to-staff ratios, ranging from 1,868 to 10,836 patients per physician FTE. On average, there are 2,018 patients registered for each FTE physician employee, with a standard deviation of 699. We estimate the following model:

$$\widehat{u}_g = \delta_0 + \delta_P PpFTE_g + \delta_C C_g + \epsilon_g^U, \tag{5}$$

where C_g includes PCP-level control variables averaged over the study period.

Results in Table 4, Column (2) show that for every additional 1,000 patients per FTE physician, $\widehat{u}_g$ increases by 0.008 ($p<0.01$). This translates to a 3.6% (95% CI [3.0%, 6.5%]) increase in the proportion of ACS attendances relative to the sample mean (0.169), or approximately 9 additional ACS attendances annually per ED for an average PCP.

We acknowledge that this analysis is not causal, as PCPs make staffing decisions endogenously. PCPs operating with higher patient-to-staff ratios may have implemented other measures (e.g., technology, auxiliary staff) to maintain service quality. Therefore, these results should be considered indicative and likely conservative, rather than definitive statements on the relationship between staffing and PCP performance.

Table 4 Relationship between the PCP performance measure $\widehat{u}_g$ and patient-to-staff ratio

	Performance measure based on ACS attendances (1) $\widehat{u}_g$	Performance measure based on Average ACS rates (2) $\widehat{u}_g$	Performance measure based on ACS admissions (3) $\widehat{u}_g$
PpFTE	0.008***	0.006	0.002
	(0.002)	(0.004)	(0.002)
Closest hospital	-0.000	0.003	-0.000
	(0.003)	(0.008)	(0.003)
Deprivation rank	-0.000	-0.002***	-0.000
	(0.000)	(0.001)	(0.000)
Scale	0.000	0.001	0.000
	(0.000)	(0.001)	(0.000)
Female	0.027	-0.021	-0.003
	(0.048)	(0.129)	(0.020)
Elderly	-0.161	-0.040	-0.000
	(0.136)	(0.384)	(0.112)
CMI	0.009	-0.017	-0.006
	(0.030)	(0.070)	(0.022)
Constant	Yes	Yes	Yes
Observations	79	79	79
R^2	0.242	0.135	0.029

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
Note. $\widehat{u}_g$ based on ACS attendances in Column (1), Average ACS rates in Column (2) and ACS admissions in Column (3).

4.4. Comparing measures based on average ACS rates

One of the key advantages of using a random effects model, as opposed to just using the average ACS rate, is its ability to account for systematic differences across PCPs and hospitals that are beyond the control of the PCP, for example practice size, patient case mix, distance to hospitals, and area deprivation levels.

To demonstrate this point, we estimate the raw average ACS rates by PCP and replicate the validation analysis using this measure. In particular we check how average ACS rates correlate with patient surveys (see Table 2 Column (3)-(4)), QOF scores (see Table 3 Column (3)-(4)), and PCP staffing levels (see Table 4 Column (2)). The estimated coefficients of average ACS rates are all closer to zero and not statistically significant, suggesting they are a more noisy and possibly biased performance measure compared to the measure $\widehat{u}_g$ obtained through the random effects model.

4.5. Comparing measures based on ACS admissions

Previous research often used hospital admissions of patients with ACS conditions via the ED to compare PCPs (Blunt 2013, Harrison et al. 2014, Barker et al. 2017). We argue that assessing PCP performance based on ACS attendances is superior to measures relying on ACS admissions. ACS admissions are rarer than attendances; in our sample, PCPs average 253 annual ACS ED attendances (SD: 130) but only 53 annual ACS admissions (SD: 35). Attendance-based measures can detect more frequent, less severe issues like appointment delays and lack of out-of-hours provisions.

Additionally, unlike admissions, ACS attendances are less influenced by hospital-specific factors such as occupancy, targets, and ED physicians' preferences (Galarraga et al. 2015, Freeman et al. 2021). This makes attendance-based measures more reflective of PCP service delivery.

To demonstrate these advantages, we replicated the analysis of the previous sections using ACS admissions. We calculated Adm_{gkt} as the proportion of ACS admissions to all admissions and re-estimated model (2). Results in Table 1, Column (4) show that systematic differences between PCPs account for much less variation in ACS admissions (17.9%, 95% CI [9.9%; 28.4%]) compared to ACS attendances (51.9%, 95% CI [37.8%; 65.7%]).

Moreover, the admission-based measure does not positively correlate with patient surveys (Table 2, Columns 5-6), QOF scores (Table 3, Columns 5-6), or staffing levels (Table 4, Column 3).

These findings support the conclusion that ACS attendances provide a more comprehensive and reliable metric for evaluating PCP performance than ACS admissions.

4.6. A placebo test based on random patient assignment

To further validate that our performance measure reflects genuine PCP differences rather than coincidental correlations, we conduct a placebo test. We compare the ICC obtained in §3 (51.9%) with the distribution of ICCs from a hypothetical scenario with no systematic PCP variability.

We construct this scenario by randomly re-assigning ED attendances to PCPs, maintaining the total number of ED attendances per PCP, hospital, and year as observed in the data. We then estimate model (2) with this randomized data. This process preserves each PCP's total ED attendance count but randomizes the mix of ACS and Non-ACS conditions, creating a random measure of A_{gkt}. We repeat this procedure 1,000 times. Figure 6 shows the resulting empirical ICC distribution. The mean ICC was 4.7% (SD: 0.036), with a maximum of 19.2%. Notably, none of these random draws produced an ICC equal to or higher than the 51.9% reported in §3. These results strongly suggest that the ICC obtained in our main analysis is highly unlikely to be due to chance, further supporting the validity of our performance measure in capturing genuine PCP differences.

5. Scalability of the Methodology: From Single Hospitals to National Data

The methodology described in §3 demonstrates its effectiveness using data from two geographically adjacent hospitals and 80 PCPs. This section aims to illustrate the scalability of this approach, both to smaller units (single hospitals) and to larger, potentially national-level datasets.

5.1. Consistency from One to Two Hospitals

To demonstrate the methodology's robustness as it scales, we first apply it separately to data from each of the two hospitals in our study. Following the steps outlined in §3, we estimate model (2) independently for Hospital 1 and Hospital 2 (full results are provided in Appendix §4).

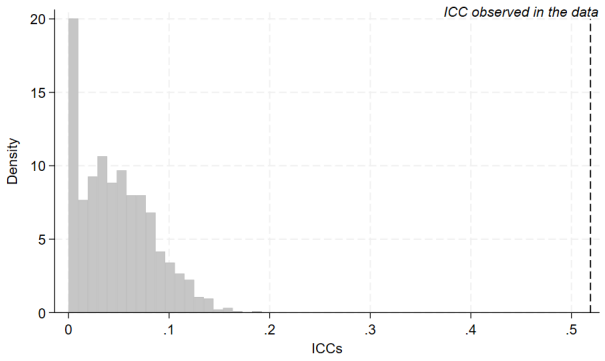

Figure 6 Empirical distribution of ICCs obtained in placebo test. The test involves 1,000 replications of assigning patients randomly to PCPs and estimating ICC using the model of (2).

The results show that there is relatively more between-PCP variability when analyzing each hospital independently, compared to the combined analysis. Specifically, the ICC estimated based on Hospital 1 is 74.1% (95% CI [48.0%; 89.9%]), and for Hospital 2 is 53.2% (95% CI [41.8%; 64.3%]). This increase in ICC is expected, as the single hospital analysis strips out any variability due to hospital specific factors. Crucially, despite these differences, the individual estimates of PCP burden on EDs ($\hat{u}_g$) are highly consistent between the single-hospital and combined analyses. Figure 8 illustrates this consistency, showing a strong positive correlation between the $\hat{u}_g$ estimates from the combined data and those from each hospital individually (correlation of 0.886 ($p<0.001$) for Hospital 1, and 0.769 ($p<0.001$) for Hospital 2).

This consistency in results when moving from single to multiple hospitals suggests that the methodology is likely to remain robust when scaled up to a larger number of hospitals, potentially at a national level.

5.2. Adapting the Methodology for National-Level Data

Scaling the methodology to national-level data, encompassing all hospitals and PCPs, requires some adaptations to the approach outlined in §3. These adaptations primarily concern the identification of influential PCPs for each hospital, the treatment of hospital-specific heterogeneity, and the refinement of hospital choice controls.

5.2.1. Identifying Influential PCPs for Each Hospital For each hospital ED $k \in \{1, ..., K\}$, it is crucial to identify the most influential PCPs. To do this in §3, we plotted the cumulative proportion of ED attendances accounted for by different PCPs and though visual inspection selected a threshold which we used to exclude low-volume PCPs. For national-level analysis, this approach can be extended by repeating this exercise for every hospital in the dataset. Alternatively, machine learning techniques such as clustering algorithms can be used to select high-volume

PCPs. As we do in §4, it would remain important to assess the robustness of the results to different threshold values.

As in the case of two hospitals, this methodology would yield a cross-classified panel structure, with PCP observations clustered in potentially more than 1 hospitals, and over time.

5.2.2. Hospital Fixed or Random Effects? In §3, we used hospital ED fixed effects (ED_k dummy variables) to control for structural differences between EDs. When scaling to national-level data with numerous hospitals, a more flexible and computationally efficient alternative, which is in line with cross-classified modeling approaches (e.g. Raudenbush 1993, Goldstein et al. 2002) is to use a random effects model, where we replace ED_k dummy variables with a random variable $\nu_k \sim N(0, \upsilon^2)$, leading to the model:

$$A_{gkt} = \alpha_0 + \alpha_C C_{gkt} + u_g + \nu_k + \varepsilon_{gkt}$$

This approach assumes that hospital effects are drawn from a Normal distribution and that hospital effects are uncorrelated with other variables in the model. It would be important to carefully evaluate the validity of these assumptions and consider alternatives if necessary.

5.2.3. Refining Hospital Choice Controls The national-level analysis allows for a more nuanced approach to controlling for hospital choice. Instead of using a binary variable ($closest_gk$) to indicate whether the study hospital was the closest option for a given PCP (as done in §3), a more sophisticated approach can be employed. This could involve using an ordinal rank variable indicating the hospital's proximity rank for each PCP, incorporating detailed geographic data to account for travel times or distances to multiple nearby hospitals, and including information on hospital specializations or service offerings that might influence patient choice beyond mere proximity. These richer controls can provide a more nuanced understanding of how patient choice impacts ED attendances and, consequently, the assessment of PCP performance.

5.3. Implementation Considerations

Scaling the proposed methodology to a national level presents both opportunities and challenges. The primary advantage lies in the potential for a comprehensive assessment of primary care performance across the entire healthcare system, offering insights into regional variations and systemic issues that could inform national policy decisions. Moreover, the increased statistical power from a larger dataset could unveil more nuanced relationships between PCP characteristics and performance, facilitating more targeted interventions to enhance healthcare delivery.

However, several challenges warrant careful consideration. Data consistency across regions is paramount, particularly in the recording and coding of ED attendances and ACS conditions. The

methodology's current focus on urban settings may not fully capture the dynamics of rural healthcare, necessitating adaptations to account for these differences. The inclusion of a larger, more diverse hospital sample may challenge the assumptions of the proposed random effects model, potentially requiring more sophisticated modeling approaches.

Computational demands will increase significantly with a larger dataset, necessitating robust data processing infrastructure. Furthermore, the potential influence of this measure on national resource allocation and reimbursement policies raises important considerations that extend beyond the scope of this study.

While this paper demonstrates the methodology's effectiveness using data from two hospitals, its applicability at a national scale remains to be empirically validated. Future research should focus on piloting this approach across diverse healthcare settings to assess its scalability and refine its implementation.

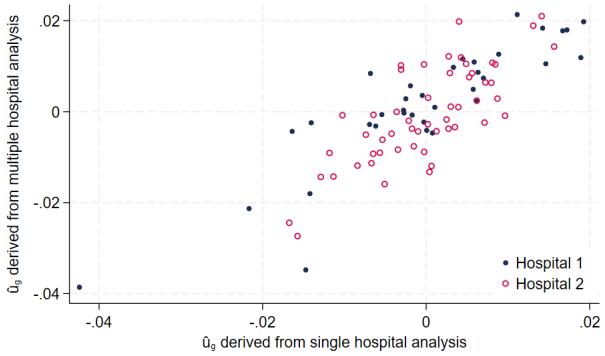

Figure 7 **Correlation between the U_g**

6. Conclusions and Managerial Implications

Key Contributions: This study introduces a novel methodology that leverages routinely collected ED operational data to assess PCP performance, offering a data-driven approach to identify and address inefficiencies in primary care delivery. By employing variance decomposition methods, we demonstrate that approximately 51.9% [95% CI: 37.8%, 65.7%] of the variation in ED attendances for Ambulatory Care Sensitive (ACS) conditions is attributable to systematic differences at the PCP level. The measure of PCP performance ($\hat{u}_g$) constructed by this methodology correlates significantly with patient satisfaction, Quality and Outcomes Framework (QOF) scores, and staffing levels, validating its relevance to service delivery. A key strength of this approach lies in its transparency, ease of implementation, and reliance on existing data, eliminating the need for additional data collection and making it readily applicable in various healthcare settings.

Practical Significance: The magnitude of these differences is operationally meaningful. The analysis suggests that if PCPs performing below the 25th percentile improved to this benchmark, it could reduce ACS attendances by 6.7% [95% CI: 4.5%, 9.4%], potentially saving the NHS £42 million annually across England. These figures are estimated using a bootstrapping method with 10,000 replications, providing robust confidence intervals for the projections (see Appendix § 5 for detailed methodology). This finding underscores the potential impact that targeted improvements in primary care could have on the broader healthcare system.

6.1. Applications and Implications

Our methodology offers several practical applications for healthcare managers and policymakers:

Resource Allocation: To demonstrate the value of the proposed measure for resource allocation, we conducted a hypothetical scenario analysis, detailed in Appendix §6. The analysis compared the impact of allocating additional physician Full-Time Equivalents (FTEs) based on different criteria: the performance measure proposed in this study ($\hat{u}_g$), patient survey results, ACS admission rates, and random assignment.

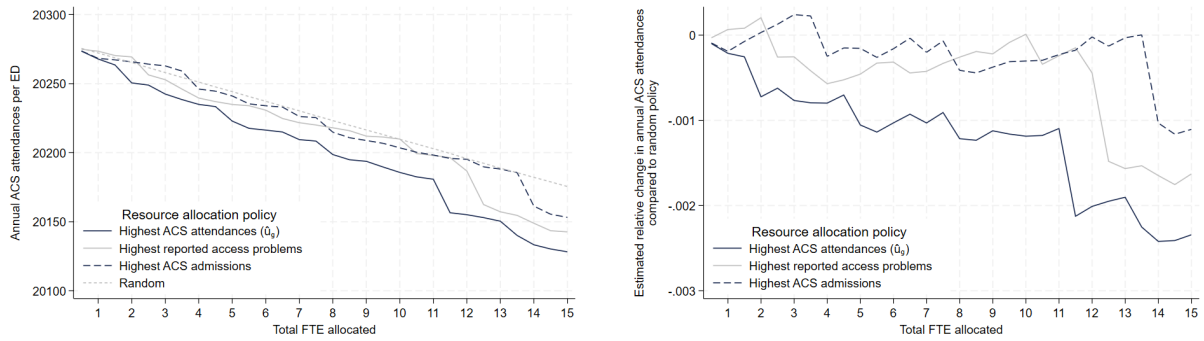

Figure 8 Impact of different resource allocation policies on ACS ED attendances.

The findings suggest that allocating resources based on the measure ($\hat{u}_g$) could lead to a more significant reduction in ACS attendances compared to other allocation strategies. Figure 9a illustrates the number of annual ACS attendances per ED for different allocation policies as the number of additional FTEs increases. Figure 9b shows the relative performance of each policy compared to random allocation. Notably, the measure ($\hat{u}_g$) outperforms other allocation strategies, including those based on patient surveys and ACS admission rates. The analysis indicates that healthcare managers might achieve more efficient ED utilization by considering this measure in resource distribution decisions.

Performance Improvement: The approach enables the identification of potential best practices and key performance drivers in primary care. The preliminary evidence suggesting a link between PCP staffing levels and performance ($\hat{u}_g$) demonstrates the potential for uncovering insights that could inform care delivery improvements. Further research incorporating more comprehensive operational and clinical data might yield additional nuanced performance drivers.

Financial Incentives: The PCP-specific metric ($\hat{u}_g$) derived from this methodology could potentially be incorporated into reimbursement formulas, such as the QOF score. This would create a form of 'yardstick competition' (Shleifer 1985), potentially aligning PCP financial incentives with system-wide efficiency goals. This approach could be implemented using existing data, offering a cost-effective avenue for performance improvement.

6.2. Limitations and Future Research

While this study assumes time-invariant PCP performance over the five-year period, future research could explore time-varying measures. Additionally, the modular nature of this approach allows for application to single hospitals or multiple hospitals, facilitating scalability to regional or national levels. Scaling to national levels would be particularly valuable as it could provide a comprehensive view of primary care performance across the entire healthcare system. This broader perspective might reveal regional variations, identify systemic issues, and inform national policy decisions. Moreover, a national-level analysis would increase the statistical power of the study, potentially uncovering more nuanced relationships between PCP characteristics and performance.

Broader Impact: This research demonstrates the potential of leveraging operational data to drive system-wide improvements in healthcare delivery. Beyond the NHS context studied here, this approach may have applications in various healthcare systems globally, including integrated care providers like Kaiser Permanente and Intermountain Healthcare, as well as entities such as the Centers for Medicare and Medicaid Services. By incorporating such data-driven approaches, both care providers and financial overseers might create more cohesive and efficient healthcare systems, potentially enhancing patient experience and outcomes.

In conclusion, this study provides a methodology for assessing PCP performance using routinely collected ED data, offering a pathway to more data-driven healthcare management. As healthcare systems worldwide face increasing demands and resource constraints, such approaches could be crucial in ensuring sustainable, high-quality care delivery.

References

Adams JL, Mehrotra A, Thomas JW, McGlynn EA (2010) Physician cost profiling—reliability and risk of misclassification. *New England Journal of Medicine* 362(11):1014–1021.

Ahmed K, Hashim S, Khankhara M, Said I, Shandakumar AT, Zaman S, Veiga A (2021) What drives general practitioners in the uk to improve the quality of care? a systematic literature review. *BMJ Open Quality* 10(1):e001127.

Ashworth M, Kordowicz M (2010) Quality and outcomes framework: smoke and mirrors? *Quality in Primary Care* 18(2):127–131.

Baker DW, Chassin MR (2017) Holding providers accountable for health care outcomes. *Annals of internal medicine* 167(6):418–423.

Barker I, Steventon A, Deeny SR (2017) Association between continuity of care in general practice and hospital admissions for ambulatory care sensitive conditions: cross sectional study of routinely collected, person level data. *The BMJ* 356:j84.

Bastani H, Goh J, Bayati M (2019) Evidence of upcoding in pay-for-performance programs. *Management Science* 65(3):1042–1060.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

Bavafa H, Canamucio A, Marcus S, Terwiesch C, Werner RM (2022) Capacity rationing in primary care: Provider availability shocks and channel diversion. *Management Science* 68(4):2377–3174.

Bell A, Fairbrother M, Jones K (2019) Fixed and random effects models: making an informed choice. *Quality & quantity* 53:1051–1074.

Berchet C (2015) Emergency care services. *OECD Health Working Papers, OECD Publishing, Paris.* (83), URL `https://www.oecd-ilibrary.org/content/paper/5jrts344crns-en`, last accessed: 30.08.2023.

Berry Jaeker J, Tucker A (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* 63(4):1042–1062.

Blunt I (2013) Focus on preventable admissions. *London: Nuffield Trust* URL `https://www.health.org.uk/sites/default/files/QualityWatch_FocusOnPreventableAdmissions.pdf`, last accessed: 05.10.2018.

Braithwaite RS (2018) Risk adjustment for quality measures is neither binary nor mandatory. *Jama* 319(20):2077–2078.

Brown C, Lilford R (2006) Cross sectional sudy of performance indicators for english PCT's: Testing construct validity and identifying explanatory variables. *BMC Health Services Research* 6:81.

Busby J, Purdy S, Hollingworth W (2015) A systematic review of the magnitude and cause of geographic variation in unplanned hospital admission rates and length of stay for ambulatory care sensitive conditions. *BMC Health Services Research* 15(1):324.

Busby J, Purdy S, Hollingworth W (2017) How do population, general practice and hospital factors influence ambulatory care sensitive admissions: A cross sectional study. *BMC Family Practice* 18(1):1–9.

Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M (2009) Effects of pay for performance on the quality of primary care in england. *New England Journal of Medicine* 361(4):368–378.

Chang CH, Stukel TA, Flood AB, Goodman DC (2011) Primary care physician workforce and medicare beneficiaries' health outcomes. *JAMA* 305(20):2096–2104.

Chen C, Savva N (2018) Unintended consequences of hospital regulation: The case of the hospital readmissions reduction program. *Available at SSRN 3236983* .

Clarke GM, Conti S, Wolters AT, Steventon A (2019) Evaluating the impact of healthcare interventions using routine data. *The BMJ* 365:l2239.

Cowling T, Harris M, Majeed A (2015) Evidence and rhetoric about access to UK primary care. *The BMJ* 350:h1513.

Cowling TE, Harris M, Majeed A (2016) Extended opening hours and patient experience of general practice in england: Multilevel regression analysis of a national patient survey. *BMJ Quality & Safety* 26:360–371.

Dolton P, Pathania V (2016) Can increased primary care access reduce demand for emergency care? Evidence from England's 7-day GP opening. *Journal of Health Economics* 49:193–208.

Dowd B, Karmarker M, Swenson T, Parashuram S, Kane R, Coulam R, Jeffery MM (2014) Emergency department utilization as a measure of physician performance. *American Journal of Medical Quality* 29(2):135–143.

Downing A, Rudge G, Cheng Y, Tu YK, Keen J, Gilthorpe MS (2007) Do the UK government's new quality and outcomes framework (QOF) scores adequately measure primary care performance? A cross-sectional survey of routine healthcare data. *BMC Health Services Research* 7(1):1–7.

Dusheiko M, Gravelle H, Jacobs R, Smith P (2006) The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment. *Journal of Health Economics* 25(3):449–478.

Dusheiko M, Gravelle H, Martin S, Rice N, Smith PC (2011) Does better disease management in primary care reduce hospital costs? Evidence from english primary care. *Journal of Health Economics* 30(5):919–932.

Filistrucchi L, Prüfer J (2019) Faithful strategies: How religion shapes nonprofit management. *Management Science* 65(1):188–208.

Freeman M, Robinson S, Scholtes S (2021) Gatekeeping under congestion: An empirical study of referral errors in the emergency department. *Management Science* 67(7):4209–4232.

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* 63(10):3147–3167.

Galarraga JE, Mutter R, Pines JM (2015) Costs associated with ambulatory care sensitive conditions across hospital-based settings. *Academic Emergency Medicine* 22(2):172–181.

Gentil ML, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, Renault E, Bouzille G, Chapron A (2017) Factors influencing the development of primary care data collection projects from electronic health records: A systematic review of the literature. *BMC Medical Informatics and Decision Making* 17(1):1–21.

Gibson OR, Segal L, McDermott RA (2013) A systematic review of evidence on the association between hospitalisation for chronic disease related ambulatory care sensitive conditions and primary health care resourcing. *BMC Health Services Research* 13(1):1–13.

Goldstein H (1994) Multilevel cross-classified models. *Sociological Methods & Research* 22(3):364–375.

Goldstein H, Browne W, Rasbash J (2002) Multilevel modelling of medical data. *Statistics in medicine* 21(21):3291–3315.

Green L, Savin S, Savva N (2013) "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* 59(10):2237–2256.

Harrison MJ, Dusheiko M, Sutton M, Gravelle H, Doran T, Roland M (2014) Effect of a national primary care pay for performance scheme on emergency hospital admissions for ambulatory care sensitive conditions: Controlled longitudinal study. *The BMJ* 349:g6423.

Hong CS, Atlas SJ, Chang Y, Subramanian S, Ashburner JM, Barry MJ, Grant RW (2010) Relationship between patient panel characteristics and primary care physician clinical performance rankings. *JAMA* 304(10):1107–1113.

Huntley A, Lasserson D, Wye L, Morris R, Checkland K, England H, Salisbury C, Purdy S (2014) Which features of primary care affect unscheduled secondary care use? A systematic review. *BMJ Open* 4(5):e004746.

Ibanez M, Toffel M (2020) How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science* 66(6):2396–2416.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Ismail SA, Gibbons DC, Gnani S (2013) Reducing inappropriate accident and emergency department attendances: A systematic review of primary care service interventions. *British Journal of General Practice* 63(617):e813–e820.

Johnson PJ, Ghildayal N, Ward AC, Westgard BC, Boland LL, Hokanson JS (2012) Disparities in potentially avoidable emergency department (ED) care: ED visits for ambulatory care sensitive conditions. *Medical Care* 50(12):1020–1028.

Jones HE, Spiegelhalter DJ (2011) The identification of "unusual" health-care providers from a hierarchical model. *The American Statistician* 65(3):154–163.

Jürges H, Köberlein J (2015) What explains drg upcoding in neonatology? The roles of financial incentives and infant health. *Journal of Health Economics* 43:13–26.

Kane RL, Shamliyan TA, Mueller C, Duval S, Wilt TJ (2007) The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care* 45(12):1195–1204.

KC D, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

KC D, Terwiesch C (2017) Benefits of surgical smoothing and spare capacity: An econometric analysis of patient flow. *Production and Operations Management* 26(9):1663–1684.

Kim SH, Chan CW, Olivares M, Escobar G (2015) ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* 61(1):19–38.

Kuntz L, Mennicken R, Scholtes S (2015) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* 61(4):754–771.

Lavoie JG, Wong ST, Ibrahim N, O'Neil JD, Green M, Ward A (2019) Underutilized and undertheorized: The use of hospitalization for ambulatory care sensitive conditions for assessing the extent to which primary healthcare services are meeting needs in british columbia first nation communities. *BMC Health Services Research* 19(1):1–10.

Li L, Georgiou A, Vecellio E, Eigenstetter A, Toouli G, Wilson R, Westbrook JI (2015) The effect of laboratory testing on emergency department length of stay: a multihospital longitudinal study applying a cross-classified random-effect modeling approach. *Academic Emergency Medicine* 22(1):38–46.

Lu SF, Rui H (2017) Can we trust online physician ratings? evidence from cardiac surgeons in florida. *Management Science* 63(10):3161–3181.

Maas CJ, Hox JJ (2004) Robustness issues in multilevel regression analysis. *Statistica Neerlandica* 58(2):127–137.

MacKichan F, Brangan E, Wye L, Checkland K, Lasserson D, Huntley A, Morris R, Tammes P, Salisbury C, Purdy S (2017) Why do patients seek primary medical care in emergency departments? An ethnographic exploration of access to general practice. *BMJ Open* 7(4):e013816.

Maisey S, Steel N, Marsh R, Gillam S, Fleetcroft R, Howe A (2008) Effects of payment for performance in primary care: Qualitative interview study. *Journal of Health Services Research & Policy* 13(3):133–139.

McCarthy ML, Zeger SL, Ding R, Levin SR, Desmond JS, Lee J, Aronsky D (2009) Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine* 54(4):492–503.

McCulloch CE, Neuhaus JM (2011a) Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science* 26(3):388–402.

McCulloch CE, Neuhaus JM (2011b) Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics* 67(1):270–279.

McDonald R, Roland M (2009) Pay for performance in primary care in england and california: Comparison of unintended consequences. *The Annals of Family Medicine* 7(2):121–127.

National Audit Office (2014) Out-of-hours gp services in england. *London: National Audit Office* URL `https://www.nao.org.uk/report/hours-gp-services-england-2/`, last access:30.08.2023.

Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K (2002) Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine* 346(22):1715–1722.

NHS (2016) Registering with a gp surgery outside the area you live. URL `https://www.nhs.uk/using-the-nhs/nhs-services/gps/patient-choice-of-gp-practices/`, last accessed: 05.10.2018.

NHS (2018) Quality and outcomes framework (qof), england 2017-2018. URL `https://files.digital.nhs.uk/B9/294E0E/qof-1718-rep.pdf`, last accessed: 19.05.2023.

NHS (2023) Gp patient survey. URL `https://gp-patient.co.uk/About`, last accessed: 24.07.2023.

Oster A, Bindman AB (2003) Emergency department visits for ambulatory care sensitive conditions: Insights into preventable hospitalizations. *Medical Care* 41(2):198–207.

Paddock SM (2014) Statistical benchmarks for health care provider performance assessment: a comparison of standard approaches to a hierarchical bayesian histogram-based method. *Health Services Research* 49(3):1056–1073.

Paddock SM, Adams JL, de la Guardia FH (2015) Better-than-average and worse-than-average hospitals may not significantly differ from average hospitals: an analysis of medicare hospital compare ratings. *BMJ Quality & Safety* 24(2):128–134.

Picard R (2012) GEONEAR: Stata module to find nearest neighbors using geodetic distances. *Statistical Software Components* .

Powell A, Savin S, Savva N (2012) Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management* 14(4):512–528.

Pronovost PJ, Angus DC, Dorman T, Robinson KA, Dremsizov TT, Young TL (2002) Physician staffing patterns and clinical outcomes in critically ill patients: A systematic review. *JAMA* 288(17):2151–2162.

Racz MJ, Sedransk J (2010) Bayesian and frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes. *Journal of the American Statistical Association* 105(489):48–58.

Raudenbush SW (1993) A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics* 18(4):321–349.

Raudenbush SW (2009) Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Education Finance and Policy* 4(4):468–491.

Roland M, Guthrie B (2016) Quality and outcomes framework: What have we learnt? *The BMJ* 354:i4060.

Rosano A, Loha CA, Falvo R, Van der Zee J, Ricciardi W, Guasticchi G, De Belvis AG (2013) The relationship between avoidable hospitalization and accessibility to primary care: A systematic review. *The European Journal of Public Health* 23(3):356–360.

Schoen C, Osborn R, Huynh PT, Doty M, Davis K, Zapert K, Peugh J (2004) Primary care and health system performance: Adults' experiences in five countries: Differing performance levels among countries highlight the potential for improvement and cross-national learning. *Health Affairs* 23(Suppl1):W4–487.

Scott A, Sivey P, Ouakrim DA, Willenberg L, Naccarella L, Furler J, Young D (2011) The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database of Systematic Reviews* (9).

Shleifer A (1985) A theory of yardstick competition. *The RAND Journal of Economics* 16(3):319–327.

Soltani M, Batt R, Bavafa H, Patterson B (2022) Does what happens in the ED stay in the ED? The effects of emergency department physician workload on post-ED care use. *Manufacturing & Service Operations Management* 24(6):2797–3306.

Song H, Andreyeva E, David G (2022) Time is the wisest counselor of all: The value of provider–patient engagement length in home healthcare. *Management Science* 68(1):420–441.

Song H, Tucker A, Murrell K (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12):3032–3053.

Song H, Tucker AL, Graue R, Moravick S, Yang JJ (2020) Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* 66(9):3825–3842.

Song H, Tucker AL, Murrell KL, Vinson DR (2018) Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science* 64(6):2628–2649.

Spiegelhalter DJ (2005) Funnel plots for comparing institutional performance. *Statistics in Medicine* 24(8):1185–1202.

Staats B, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management Science* 58(6):1141–1159.

Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* 60(6):1574–1593.

Thomas JW, Grazier KL, Ward K (2004) Economic profiling of primary care physicians: Consistency among risk-adjusted measures. *Health Services Research* 39(4p1):985–1004.

van Loenen T, van den Berg MJ, Westert GP, Faber MJ (2014) Organizational aspects of primary care related to avoidable hospitalization: A systematic review. *Family Practice* 31(5):502–516.

Vuik SI, Fontana G, Mayer E, Darzi A (2017) Do hospitalisations for ambulatory care sensitive conditions reflect low access to primary care? An observational cohort study of primary care usage prior to hospitalisation. *BMJ Open* 7(8):e015704.

Wang J, Zhou YP (2018) Impact of queue configuration on service time: Evidence from a supermarket. *Management Science* 64(7):3055–3075.

Watson AG, McCoy JV, Mathew J, Gundersen DA, Eisenstein RM (2019) Impact of physician workload on burnout in the emergency department. *Psychology, Health & Medicine* 24(4):414–428.

Weber EJ, Mason S, Freeman JV, Coster J (2012) Implications of england's four-hour target for quality of care and resource use in the emergency department. *Annals of Emergency Medicine* 60(6):699–706.

Weinick RM, Burns RM, Mehrotra A (2010) Many emergency department visits could be managed at urgent care centers and retail clinics. *Health Affairs* 29(9):1630–1636.

Weintraub WS, Garratt KN (2017) Challenges in risk adjustment for hospital and provider outcomes assessment. *Circulation* 135(4):317–319.

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT Press).

Xu Y, Tan T, Netessine S (2020) The impact of workload on operational risk: Evidence from a commercial bank. *SMU Cox: IT & Operations Management (Topic)* .