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

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

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

[#] These authors contributed equally to this work and should be considered co-first authors.

This online appendix presents additional investigations and robustness checks on the analysis presented in the main paper.

Contents

1. List of ACS conditions and publicly available databases used in this study

Table 1 List of ACS conditions

Acute	Chronic	Vaccine-preventable
Cellulitis	Angina	Influenza
Dehydration	Asthma	Pneumonia
Dental conditions	Chronic obstructive pulmonary disease	Tuberculosis
Ear, nose, and throat infections	Congestive heart failure	Other vaccine-preventable conditions
Gangrene	Convulsions and epilepsy	
Gastroenteritis	Diabetic complications	
Nutritional deficiencies	Hypertension	
Pelvic inflammatory disease	Iron deficiency	
Perforated or bleeding ulcer		
Urinary tract infections or pyelonephritis		

Table 2 Publicly available databases used in this study

Publisher, Data set	Extracted Information	Link
NHS Digital, Epraccur	Name and addresses of PCPs and when they ended operating (time of closure)	https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data
NHS Digital, Ebranches	PCP branches and when they started operating	https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data
NHS Digital, Patients Registered at a GP Practice	Number and demographics of patients registered at PCP	https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice
NHS Digital, Payments to General Practice	PCP patient case-mix	https://digital.nhs.uk/data-and-information/publications/statistical/nhs-payments-to-general-practice
NHS Digital, Etrust	Location of hospital sites	https://digital.nhs.uk/services/organisation-data-service/data-downloads/other-nhs-organisations
NHS Digital, General and Personal Medical Services	FTE physicians working at PCPs	https://webarchive.nationalarchives.gov.uk/20180328140045/http://digital.nhs.uk/catalogue/PUB20503
NHS Digital, Quality & Outcomes Framework	QOF achievements per PCP	https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement
NHS England, GP Patient Survey	Patient perception of PCPs	https://gp-patient.co.uk/about
Free Map Tools, UK Postcode Database	Latitude and longitude coordinates of UK postcodes	https://www.freemaptools.com/contact.htm
Ministry of Housing, Communities & Local Government	English indices of deprivation at small area level	https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015
Office for National Statistics	Postcode to Output Area	https://geoportal.statistics.gov.uk/datasets/e7824b1475604212a2325cd373946235/about

2. Descriptive statistics for the PCPs

Table 3 Descriptive statistics for the 80 PCPs.

		2013	2014	2015	2016	2017
Proportion of ACS ED visits	Mean	0.171	0.173	0.166	0.167	0.165
	SD	0.037	0.036	0.035	0.041	0.029
Proportion of PCPs closest to EDs	Mean	0.148	0.148	0.148	0.140	0.140
	SD	0.357	0.357	0.357	0.349	0.349
Deprivation rank (in 1,000)	Mean	9.031	9.031	9.031	9.159	9.159
	SD	5.220	5.220	5.220	5.211	5.211
PCP scale (in 1,000 patients)	Mean	8.455	8.757	9.253	8.998	9.447
	SD	3.872	3.936	3.898	3.653	3.794
Proportion of female patients	Mean	0.505	0.495	0.499	0.495	0.495
	SD	0.043	0.043	0.021	0.042	0.041
Proportion of elderly patients	Mean	0.035	0.034	0.035	0.034	0.033
	SD	0.013	0.012	0.012	0.013	0.013
Case-mix Index	Mean	0.998	0.984	0.983	0.971	0.967
	SD	0.069	0.067	0.065	0.063	0.063

3. Detailed validation results

In section §4.1-4.3 of the main paper, we validate the PCP performance measure $\widehat{u}_g$ by examining whether it is correlated with patient survey outcomes, QOF scores, and staffing decisions made by the PCPs. This section replicates the validation exercise for the alternative measures discussed in sections §4.4-4.5 of the main paper and compares them to the corresponding validation results. (Columns (1)–(2) of Tables 4 and 5, and Column (1) of Table 6, repeat the validation results from the main paper.)

We first consider the validation based on patient survey outcomes (Table 4). For $\widehat{u}_g$, the coefficients in Columns (1)–(2) are positive and statistically significant. Replacing $\widehat{u}_g$ with raw average ACS rates (Columns (3)–(4)) or ACS admissions (Columns (5)–(6)) yields coefficients that are closer to zero and statistically insignificant, indicating that neither alternative measure reproduces the survey-based validation results.

We next turn to the validation based on QOF scores (Table 5). While $\widehat{u}_g$ exhibits a significant association with high QOF achievement in Column (1), neither average ACS rates nor ACS admissions produce statistically significant coefficients in Columns (3)–(6), suggesting that these alternative measures do not capture systematic differences in clinical performance reflected in QOF scores.

Finally, we examine the validation based on staffing decisions (Table 6). $\widehat{u}_g$ is positively and significantly associated with patient-to-staff ratios in Column (1), whereas the corresponding coefficients for average ACS rates (Column (2)) and ACS admissions (Column (3)) are again closer to zero and statistically insignificant.

Taken together, these results indicate that, unlike $\widehat{u}_g$, neither raw average ACS rates nor ACS admissions yield consistent validation across patient surveys, QOF scores, and staffing decisions, reinforcing the conclusion that ACS attendances provide a more comprehensive and reliable metric for evaluating PCP performance.

Table 4 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 FE	Yes	Yes	Yes	Yes	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 $> \chi^2$	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).

Table 5 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) $P(QOF_{gt} \geq 90\%)$	(2) $P(QOF_{gt} \geq 95\%)$	(3) $P(QOF_{gt} \geq 90\%)$	(4) $P(QOF_{gt} \geq 95\%)$	(5) $P(QOF_{gt} \geq 90\%)$	(6) $P(QOF_{gt} \geq 95\%)$
$\widehat{u}_g$	-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 FE	Yes	Yes	Yes	Yes	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 $> \chi^2$	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).

Table 6 Relationship between the PCP performance measure $\widehat{u}_g$ and patient-to-staff ratio

	Performance measure based on ACS attendances	Performance measure based on Average ACS rates	Performance measure based on ACS admissions
	(1) $\widehat{u}_g$	(2) $\widehat{u}_g$	(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. Alternative model specifications

For the model presented in (1),

$$A_{gkt} = \alpha_0 + \alpha_C C_{gkt} + u_g + \epsilon_{gkt}, \tag{1}$$

we normalize the number of ACS attendances by dividing with the total number of attendances. This normalization controls for heterogeneity that affects patients with ACS and Non-ACS conditions equally. In addition, we explicitly control for observed PCP and ED heterogeneity that might affect ACS patients differently than Non-ACS patients. In this section we investigate three alternative modeling choices.

4.1. Adaptive centering approach

In this section 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.

To obtain the transformed variables, we follow the three-step procedure outlined by Raudenbush (2009, p. 489): First, we regress each covariate ($Scale, Female$, etc.) on the PCP identifier $g \in 1, ..., G$ and save the residuals. Next, we construct indicator variables Z_{gkt}, where $Z_{gkt} = 1$ if the unit-of-analysis for PCP g was recorded at hospital k in year t, and $Z_{gkt} = 0$ otherwise. These indicators are created for each hospital-year combination ($k \in 1, 2$, $t \in 1, .., 5$), resulting in 10 dummy variables. Each dummy variable is then regressed on the PCP identifier g and we save the residuals. Finally, for each covariate ($Scale, Female$,etc.), we regress the Step 1 residuals on the 10 residuals from Step 2. The residuals from this regression serve as the transformed covariates used in the subsequent analysis.

We estimate the model in (1) using these transformed variables for time-varying PCP covariates (results reported in Table 7, Column (2)) and compare the resulting $\widehat{u}_g$ with those from the primary model (for comparison, Table 7, Column (1) repeats the results of the main paper). The high correlation (0.920, p<0.001) between these estimates suggests robustness to time-invariant confounding from observable PCP covariates. As shown in Figure 1, the histograms and kernel density estimates for $\widehat{u}_g$ from both models suggest that there are no significant differences between the distributions generated by these two methods. This is further supported by a Kolmogorov-Smirnov test for equality of distributions, where the p-value is 1.000. 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. Therefore, we also estimate a fixed-effects model in the next section.

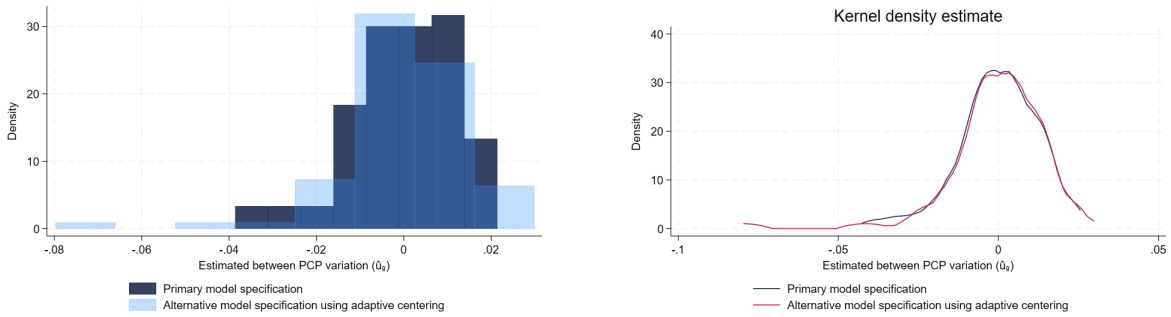

Figure 1 $\widehat{u}_g$ **comparison primary model vs. adaptive centering**

4.2. Fixed-effect specification

In this section we estimate model (1) by treating the unobservable heterogeneity between PCPs u_g as a fixed effect. In contrast to the random-effects model estimated in the main paper, this specification does not place any parametric assumptions on PCP heterogeneity (the random-effects model assumed Normally distributed heterogeneity) and does not assume that PCP heterogeneity is orthogonal to any of the time varying controls included in the model. We note that this model specification cannot include the distance variable $closest_{gk}$ or the deprivation index D_g as they are time-invariant and are therefore collinear with the PCP fixed effects. We estimate the fixed-effect model using the Stata command *xtreg, fe* and provide the results in Table (7) Column (4). In the fixed-effect specification 77.5% of the variation is attributed to systematic differences between PCPs compared to the 51.9% in the random-effect specification (Table (7) Column (1)). The larger between-variation in the fixed-effect specification may be the result of absorbing time-invariant heterogeneity between PCPs that is explicitly controlled for in the random-effect specification.

We argued in the main paper that due to the low number of observations per PCP the fixed-effect specification will estimate the unobservable heterogeneity between PCPs with more uncertainty than the random-effect specification. This is illustrated in Figure 2. In fact, the variance of $\widehat{u}_g$ is 0.00054122 in the fixed-effect specification, which is 214% higher than the variance of the random-effect specification (0.0001722).

We then compare PCP performance rankings under random-effects (RE) and fixed-effect specification (FE). We find that while the two models produce highly correlated rankings (correlation=0.751, p<0.001), there are systematic and meaningful differences that indicate the random-effect model is, in fact, adjusting for time-invariant factors rather than simply mirroring fixed-effect estimates.

The left panel of Figure 3 below illustrates this comparison. While rankings between the models generally align, some PCPs experience notable shifts: PCP A and PCP B, for example, rank 50$^{\text{th}}$ and 52$^{\text{nd}}$ under FE but move up to 18$^{\text{th}}$ and 19$^{\text{th}}$ under RE. Both practices operate in highly

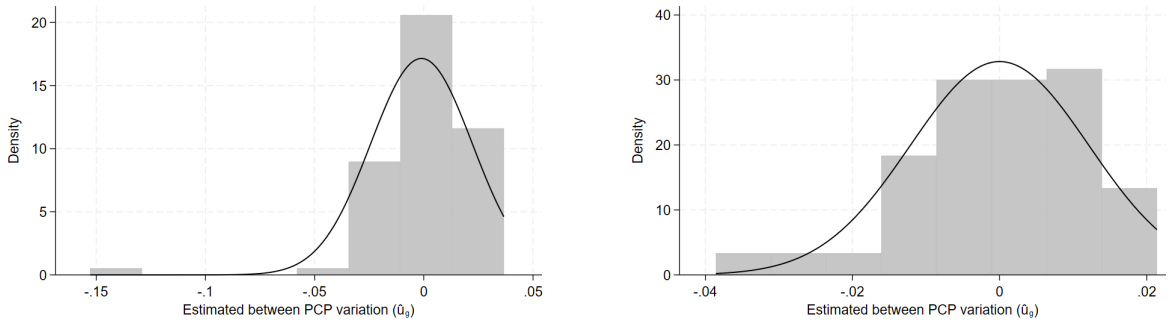

Figure 2 **Variation between PCPs: Fixed-effect specification (left) and Random-effect specification (right)**

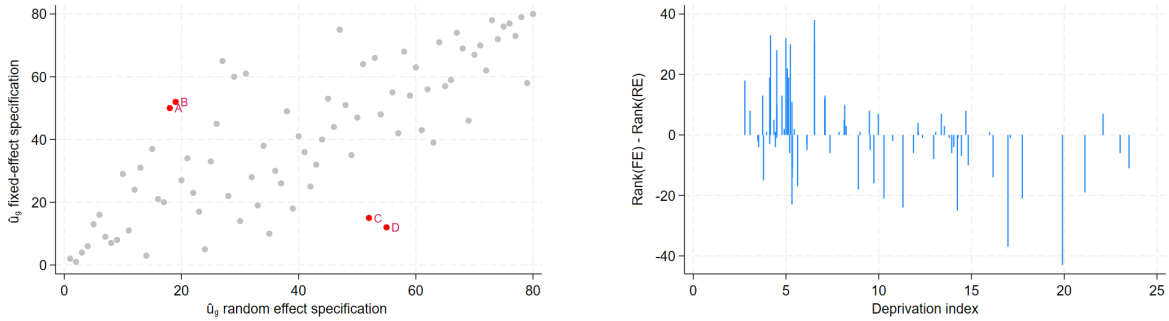

Figure 3 **PCP performance rankings under random-effects and fixed-effect specification**

deprived areas (deprivation indices = 4.9 and 4.2, far more deprived than the sample average of 9.0). The FE model absorbs these deprivation effects, leading to a lower ranking, whereas the RE model explicitly adjusts for these structural challenges. Conversely, PCP C and PCP D rank 15$^{\text{th}}$ and 12$^{\text{th}}$ under FE but drop to 52$^{\text{nd}}$ and 55$^{\text{th}}$ under RE. These PCPs are in affluent areas (deprivation indices = 16.9 and 19.9). Since the FE model does not adjust for deprivation effects, these PCPs appear to perform better under FE than under RE.

The right panel of Figure 3 further confirms this pattern: PCPs in more deprived areas systematically rank better under RE than FE (Rank (FE) – Rank (RE) > 0). PCPs in affluent areas systematically rank worse under RE than FE (Rank (FE) – Rank (RE) < 0).

Thus, while the overall correlation between FE and RE is strong, the key takeaway is that the RE model is effectively adjusting for structural differences, ensuring a more equitable comparison of PCP performance.

For validation purposes, we re-estimate the different models outlined in §4 of the main paper, i.e. first a model using the results of the patient survey:

$$P_{gt}^{i} = \beta_0 + \beta_U \widehat{u}_g + \beta_\epsilon \widehat{\epsilon}_{gt} + \beta_C C_{gt} + u_g^P + \epsilon_{gt}^P, \tag{2}$$

second, a model using the QOF scores:

$$\begin{aligned} HighQ^*_{gt} &= \gamma_0 + \gamma_U \widehat{u}_g + \gamma_\epsilon \widehat{\epsilon}_{gt} + \gamma_C C_{gt} + u^Q_g + \epsilon^Q_{gt}, \\ HighQ_{gt} &= 1[HighQ^*_{gt} > 0], \end{aligned} \tag{3}$$

and third, a model using the patient-to-staff ratio as a performance antecedent:

$$\widehat{u}_g = \delta_0 + \delta_P PpFTE_g + \delta_C C_g + \epsilon^U_g. \tag{4}$$

The results of estimating models (2) – (4) with $\widehat{u}_g$ obtained through the fixed-effect specification appear in Table 8 – 10 and are similar to the results reported in the main paper. We find that PCPs with a higher $\widehat{u}_g$ are also the ones patients are more likely dissatisfied with, PCPs with a higher $\widehat{u}_g$ are associated with a lower probability of obtaining QOF scores of at least 90%, and we find positive association between patient-to-staffing ratio and $\widehat{u}_g$.

4.3. Testing non-linear scale effects

One of the controls included in the model of (1) is PCP scale. This is the number of patients registered with a PCP. In this section we test whether scale might affect PCP performance in a non-linear fashion by adding $Scale_{gt} \times Scale_{gt}$ to the model. For comparison, Column (1) in Table 7 repeats the results of the main paper and Column (3) shows the results of the model with $Scale_{gt} \times Scale_{gt}$. There is no evidence of non-linear scale effects as the coefficients of $Scale_{gt}$ and $Scale_{gt} \times Scale_{gt}$ are close to zero (and not statistically significant).

Table 7 Decomposing variation in PCP performance: Alternative model specifications

	(1) A_{gkt}	(2) A_{gkt}	(3) A_{gkt}	(4) A_{gkt}	(5) A_{gkt}
Closest hospital	-0.002	-0.000	-0.002		-0.001
	(0.003)	(0.003)	(0.003)		(0.004)
Deprivation rank	-0.001***	-0.001***	-0.001***		
	(0.000)	(0.000)	(0.000)		
Scale	0.000	0.001	-0.002	0.001	0.000
	(0.000)	(0.001)	(0.002)	(0.001)	(0.000)
Scale × Scale			0.000		
			(0.000)		
Female	0.137	-0.233	0.156	-0.233	0.107
	(0.101)	(0.143)	(0.103)	(0.145)	(0.120)
Elderly	-0.170	0.330	-0.184	0.330	-0.255
	(0.161)	(0.375)	(0.159)	(0.381)	(0.156)
CMI	0.061*	0.040	0.056*	0.040	0.101***
	(0.033)	(0.051)	(0.034)	(0.051)	(0.029)
Constant	Yes	Yes	Yes	Yes	Yes
ED × Year FE	Yes	Yes	Yes	Yes	Yes
PCP Effect	Random	Random	Random	Fixed	Random
$\widehat{\tau}^2$	0.0001722	0.0002492	0.0001657	0.00054122	0.0002039
$\widehat{\tau}^2$ 95% CI	[.0001038; .0002856]	[.0001083; .0005735]	[.0000981; .0002801]	n/a	[.0001219; .0003411]
ICC	51.9%	62.1%	50.9%	77.5%	56.0%
ICC 95% CI	[37.8%; 65.7%]	[41.0%; 79.5%]	[36.3%; 65.2%]	n/a	[41.7%; 69.5%]
Wald χ^2 (F-statistic)#	916.5	915.2	1,008.4	33.8	825.3
Prop $> \chi^2$	0.000	0.000	0.000	0.000	0.000
Log Pseudolikelihood	1,182.1	1,177.9	1,182.8	n/a	1,175.8
Observations	426	426	426	426	426
Number of groups	80	80	80	80	80

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
Note. # F-statistic applies to column (4). na: statistics not provided in the fixed-effects estimation.

Table 8 Estimating model (2) using $\widehat{u}_g$ obtained from a fixed-effect specification

	(1) Not rec	(2) No access
$\widehat{u}_g$	0.914***	0.886***
	(0.334)	(0.289)
$\widehat{\epsilon}_{gt}$	-0.053	0.501**
	(0.222)	(0.246)
Closest hospital	-0.008	-0.005
	(0.015)	(0.014)
Deprivation rank	0.001	-0.002***
	(0.001)	(0.001)
Scale	-0.001	0.001
	(0.001)	(0.001)
Female	-0.414**	0.001
	(0.190)	(0.183)
Elderly	1.865***	1.699***
	(0.527)	(0.496)
CMI	-0.246**	-0.314***
	(0.108)	(0.087)
Constant	Yes	Yes
Year FE	Yes	Yes
PCP effect	Random	Random
Wald χ^2	21.0	59.7
Prop $> \chi^2$	0.051	0.000
Log Pseudolikelihood	614.7	607.4
Observations	387	387
Number of groups	80	80

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1
Not rec: Proportion of patients refraining to recommend their PCP. No access: Proportion of patients experiencing access problems at their PCP.

Table 9 **Estimating model (3) using $\widehat{u}_g$ obtained from a fixed-effect specification**

	(1) $P(QOF_{gt} \geq 90\%)$	(2) $P(QOF_{gt} \geq 95\%)$
$\widehat{u}_g$	-21.025*	6.380
	(11.344)	(13.849)
$\widehat{\epsilon}_{gt}$	11.589	15.960*
	(12.074)	(8.723)
Closest hospital	0.899	0.795
	(0.687)	(0.550)
Deprivation rank	-0.056	-0.010
	(0.043)	(0.049)
Scale	-0.038	-0.027
	(0.054)	(0.067)
Female	12.350*	4.465
	(6.321)	(7.459)
Elderly	-28.097	14.372
	(19.494)	(19.254)
CMI	-1.128	-5.220
	(3.840)	(3.867)
Constant	Yes	Yes
Year FE	Yes	Yes
PCP effect	Random	Random
Wald χ^2	25.6	24.7
Prop $> \chi^2$	0.012	0.016
Log Pseudolikelihood	-115.6	-202.7
Observations	389	389
Number of groups	80	80

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 10 **Estimating model (4) using $\widehat{u}_g$ obtained from a fixed-effect specification**

	(1) $\widehat{u}_g$
PpFTE	0.009***
	(0.002)
Closest hospital	-0.002
	(0.003)
Deprivation rank	-0.001***
	(0.000)
Scale	-0.001**
	(0.000)
Female	0.397***
	(0.061)
Elderly	-0.667***
	(0.165)
CMI	0.025
	(0.036)
Constant	Yes
Observations	79
R^2	0.692

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

5. Exploring the Impact of Correlation between Deprivation and PCP effects

This section provides detailed results for the robustness analyses that examine the sensitivity of our findings to alternative treatments of socioeconomic deprivation. Building on the discussion in the main paper, we outline our simulation approach (§5.1) and report the full results from the simulation analysis (§5.2), risk-stratified specifications (§5.3), and alternative model formulations (§5.4) used to assess the implications of potential correlation between deprivation and PCP performance.

5.1. Simulation approach

To evaluate how potential correlations between u_g (PCP performance) and economic deprivation could impact our results, we construct counterfactual scenarios. Specifically, we hold the empirical distribution of the deprivation index (D_g) constant while systematically varying its correlation with u_g. One challenge arises from the highly skewed, non-normal distribution of D_g (illustrated in Figure 4). To address this, we apply the Gaussian copula method, which effectively generates correlated variables without altering the original marginal distributions of D_g and u_g. This method ensures that simulated scenarios remain realistic and consistent with our observed data. However, it is important to note that, given the fixed marginal distributions of D_g and u_g, only certain correlation values are mathematically feasible. Therefore, our analysis explores correlations within the practically achievable range determined by these inherent distributional constraints.

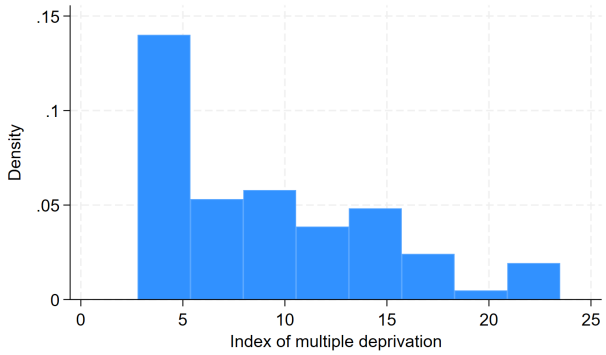

Figure 4 Distribution of the index of multiple deprivation (D_g)

The simulation approach consists of the following steps:

1. *Establish Baseline Estimates.* We begin by estimating the baseline model using the random-effects specification described earlier in the paper. This step relies on the standard assumptions underlying the random-effects framework – specifically, that the PCP performance metric (u_g) is normally distributed and independent of the observed covariates included in our model. Under these assumptions, we estimate the baseline performance decomposition model as follows:

$$A_{gkt} = \alpha_0 + \alpha_C C_{gkt} + \alpha_D D_g + u_g + \epsilon_{gkt}, \tag{5}$$

where A_{gkt} denotes the proportion of ACS attendances for PCP g at hospital k in year t, C_{gkt} represents a vector of PCP-specific characteristics, and D_g is the time-invariant deprivation index. After estimating this model, we extract the empirical distributions of the PCP-specific random effects ($\widehat{u}_g$) and residuals ($\widehat{\epsilon}_{gkt}$). Additionally, we explicitly obtain and store the estimated coefficients: $\widehat{\alpha}_0, \widehat{\alpha}_C, \widehat{\alpha}_D$. These coefficients serve as baseline parameters for constructing the counterfactual scenarios in subsequent simulation steps.

2. *Transforming $\widehat{u}_g$ to normal space.* Because the estimated PCP random effects ($\widehat{u}_g$) follow an empirical distribution that may deviate from normality (see Figure 2b), we follow the usual practice of transforming them into a standard normal distribution. We accomplish this transformation by first converting each PCP effect estimate ($\widehat{u}_g$) into percentiles ranging from 0 to 1 based on their empirical distribution. We then map these percentiles onto corresponding quantiles of a standard normal distribution. The resulting transformed variable (nu_g) retains the original rank ordering of the PCP estimates while conforming to the normality assumption essential for implementing the Gaussian copula approach.

3. *Generating a Correlated Normal Variable.* To systematically introduce and control the correlation between the transformed PCP performance measure (nu_g) and the deprivation index, we generate a new variable (z_1) defined as:

$$z_1 = \rho\, nu_g + \sqrt{1-\rho^2}\, z_2, \tag{6}$$

where $z_2 \sim \mathcal{N}(0,1)$ is an independent standard normal variable, and ρ is the correlation coefficient between z_1 and nu_g. By construction, z_1 is also standard normal, with a mean of 0 and variance of 1. We draw random values of ρ from a uniform distribution in the range of [-0.9; 0.9]. Every draw represents a different scenario with a different magnitude and direction of correlation.

4. *Generating a Counterfactual Deprivation Index.* Next, we use the newly generated correlated normal variable (z_1) to construct a counterfactual deprivation index ($\tilde{D}_g$). To ensure realism, $\tilde{D}_g$ must preserve the exact empirical distribution (including skewness, range, and shape) of the original deprivation index (D_g). We do this by applying a rank-based mapping procedure: we sort the values of z_1 and match them to the sorted empirical values of the original deprivation index D_g. This procedure results in a new deprivation index, $\tilde{D}_g$, that precisely retains the original marginal distribution of D_g while enforcing the controlled correlation structure between deprivation and PCP performance ($\widehat{u}_g$).

5. *Computing Counterfactual Outcomes.* Using the counterfactual deprivation index ($\tilde{D}_g$), we compute the counterfactual proportion of ACS attendances ($\tilde{A}_{gkt}$). Specifically, we apply the previously estimated model coefficients and the original empirical distributions of the PCP random effects ($\widehat{u}_g$) and residuals ($\widehat{\epsilon}_{gkt}$) as follows:

$$\tilde{A}_{gkt} = \widehat{\alpha}_0 + \widehat{\alpha}_C C_{gkt} + \widehat{\alpha}_D \tilde{D}_g + \widehat{u}_g + \widehat{\epsilon}_{gkt}. \tag{7}$$

This step generates realistic, model-consistent predictions of how ACS attendance proportions would differ under the specified counterfactual correlation between PCP performance and socioeconomic deprivation.

6. *Re-estimating the Model and Assessing Impact.* We re-estimate the performance decomposition model using the newly computed counterfactual ACS attendance proportions ($\tilde{A}_{gkt}$). Based on this re-estimation, we calculate the Intraclass Correlation Coefficient (ICC) to quantify how variations in PCP performance change under the introduced correlations. Additionally, we derive updated estimates of the PCP random effects ($\tilde{u}_g$) and re-rank the PCPs accordingly. To evaluate the robustness of our findings, we compute the rank correlation between these new PCP rankings and the original baseline rankings. This rank correlation comparison allows us to determine the stability of PCP performance rankings under different correlation scenarios.

7. *Accounting for Statistical Variability.* We repeat steps 3 through 6 for a total of 1,000 scenarios.

5.2. Simulation results

By construction, the Gaussian copula method used to simulate different correlation scenarios between PCP performance (u_g) and the deprivation index (D_g), preserves the exact marginal distributions of the original variables while allowing the correlation between them to vary. We show the correlations r_s generated by this method in Figure 5. Note that the correlations produced by the simulation are in the range of -0.64 to $+0.91$. Correlations outside this range are not possible because of the intrinsic properties of the marginal distributions. Specifically, given the empirical shapes (such as skewness, kurtosis, and discreteness) of u_g and D_g, there are inherent mathematical constraints on the achievable correlation values. In other words, certain combinations of marginal distributions inherently restrict correlations from reaching values close to ± 1.

We summarize the impact of varying correlation levels (r_s) between PCP performance (u_g) and deprivation (D_g) on two important metrics: the Intraclass Correlation Coefficient (ICC), representing variability explained by PCP differences, and Spearman's rank correlation, measuring stability in PCP rankings. Figure 6 shows these results; key findings are:

- *"Zero" correlation ($-0.05 \leq r_s \leq 0.05$):*

  — All ICC values fall within the range of [50.7%; 52.1%] and closely align with the original ICC estimate of 51.9%, confirming internal consistency.

  — The rank correlation between PCP rankings remains stable, all ranging from 0.98 to 1, as expected, since this scenario matches the baseline assumption.

- *Weak to moderate correlation ($-0.5 \leq r_s < -0.05$ or $0.05 < r_s \leq 0.5$):*

  — ICC slightly decreases, ranging from 38.3% to 52.1%, indicating mild underestimation of PCP variation. However, deviations remain small.

—PCP rankings remain highly stable, with rank correlations ranging from 0.79 to 1 in 95% of the cases, indicating robustness in identifying top- and bottom-performing PCPs under moderate correlation scenarios.

- *Strong negative correlation ($r_s < -0.5$):*

—ICC declines more substantially, dropping to between 33.7% and 46.7%, suggesting a stronger bias due to the violation of independence assumptions.

—PCP rankings become moderately less stable but remain relatively consistent, with 95% of correlations falling between 0.71 and 0.86.

- *Strong positive correlations ($r_s > 0.5$):*

—ICC values sharply decline, ranging from 14.7% to 45.8%, demonstrating significant bias in PCP performance estimation.

—PCP ranking stability deteriorates notably, with rank correlations dropping substantially to between 0.42 and 0.84 in 95% of cases. This indicates meaningful deviations in identifying PCP performance under strong positive correlations.

Overall, these findings confirm theoretical expectations: the random-effects model becomes inconsistent if the random effects are strongly correlated with covariates such as socio-economic deprivation. Nevertheless, at realistic (mild-to-moderate) correlation levels, both the ICC and PCP rankings remain robust and stable. Moreover, the practically relevant scenario of negative correlation – where PCPs in more deprived areas tend to perform worse – is characterized by relatively minor biases, reinforcing the empirical validity and robustness of our main findings.

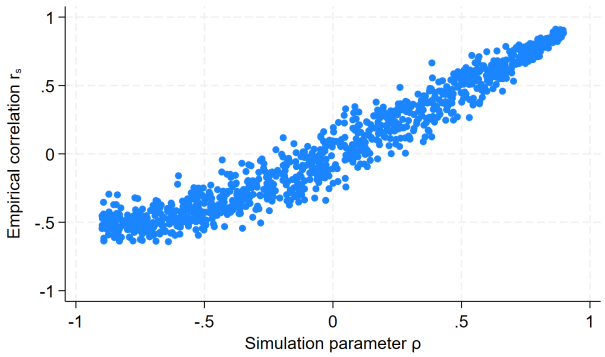

Figure 5 Empirical correlation r_s under varying levels of simulation parameter ρ

5.3. Risk-stratification

Next, we apply risk stratification by comparing PCPs within peer groups that face similar socioeconomic challenges. We form these peer groups using the deprivation index, dividing PCPs into

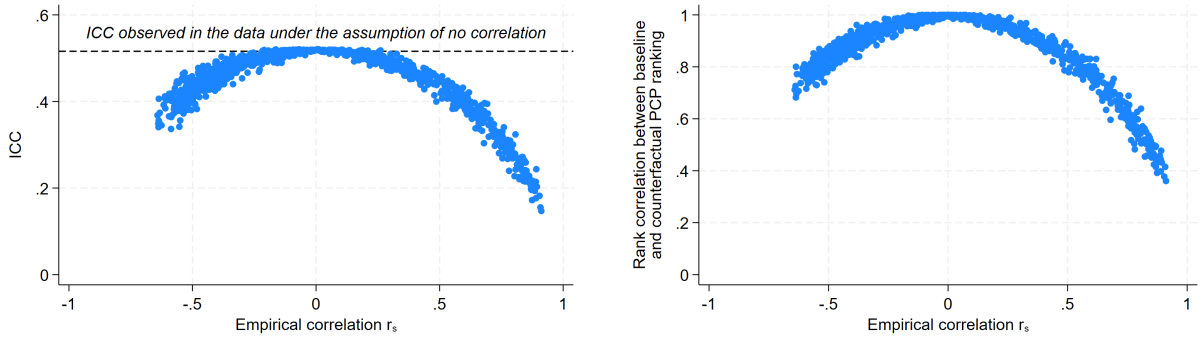

Figure 6 Estimated ICC (a) and Rank correlation between the baseline PCP performance ranking and the counterfactual rankings (b) under varying levels of correlation r_s

terciles, which results in 26–27 PCPs and 134–155 observations per tercile. While finer stratifications (e.g., quintiles) would increase socioeconomic homogeneity, they would substantially reduce group sizes and undermine estimation reliability in our setting. Given this trade-off and the size of our dataset, deprivation terciles provide a balanced and feasible approach to defining comparable peer groups.

Within each deprivation tercile, we replicate the performance decomposition model (Table 11) and compute PCP-specific performance measures and rankings. Across all terciles, the ICCs are lower than in the pooled analysis (51.9%), with the lowest ICC (34.8%) observed among PCPs operating in the most deprived areas. This pattern indicates that, within more deprived contexts, a smaller share of outcome variation is attributable to systematic differences between PCPs, consistent with more limited scope for differentiation among practices serving similar populations. As expected given the reduced sample sizes within terciles, these estimates are accompanied by wider confidence intervals for both the between-PCP variance component and the ICC.

Using the performance metric $\widehat{u}_g$ from the risk-stratified analyses, we compute PCP performance rankings and compare them with those from the baseline specification. Figure 7 displays these comparisons. The resulting rank correlations are substantial across all groups, ranging from 0.650 in the most deprived tercile to 0.861 in the least deprived tercile. These magnitudes indicate meaningful alignment between stratified and pooled rankings and demonstrate that high- and low-performing practices can still be distinguished when comparisons are restricted to socioeconomic peer groups, although the degree of alignment is stronger in less deprived settings.

5.4. Alternative specification – Omitting deprivation

As a final step to assess the role of deprivation in our results, we estimate an alternative specification that omits the deprivation index from the set of covariates. Column (5) of Table 7 reports the results. Relative to the primary analysis (Column (1)), a larger share of the variation is attributed

Table 11 Risk-Stratified Analyses per Deprivation Tercile

Variables	1st tercile	2nd tercile	3rd tercile
Closest hospital	-0.008**	-0.003	0.023***
	(0.004)	(0.004)	(0.007)
Scale	0.001	-0.001**	0.001*
	(0.000)	(0.001)	(0.001)
Female	-0.394***	-0.022	0.184***
	(0.105)	(0.120)	(0.049)
Elderly	-0.524*	-0.016	-0.094
	(0.301)	(0.206)	(0.228)
CMI	0.149***	0.016	0.021
	(0.051)	(0.038)	(0.048)
Constant	Yes	Yes	Yes
ED × Year FE	Yes	Yes	Yes
$\widehat{\tau}^2$	0.0000842	0.0001067	0.0001503
$\widehat{\tau}^2$ 95% CI	[0.0000404; 0.0001755]	[0.0000521; 0.0002186]	[0.0000610; 0.0003704]
ICC	34.8%	45.8%	49.4%
ICC 95% CI	[18.5%; 55.6%]	[27.2%; 65.7%]	[25.8%; 73.2%]
Wald χ^2	1,746.7	728.0	717.8
Prop $> \chi^2$	0.000	0.000	0.000
Log Pseudolikelihood	439.8	389.1	383.5
Spearman's ρ #	0.650	0.815	0.861
Observations	155	134	137
Number of groups	27	27	26

*** p<0.01, ** p<0.05, * p<0.1.
# Rank correlation between provider rankings from the primary analysis and the risk-stratified analyses.

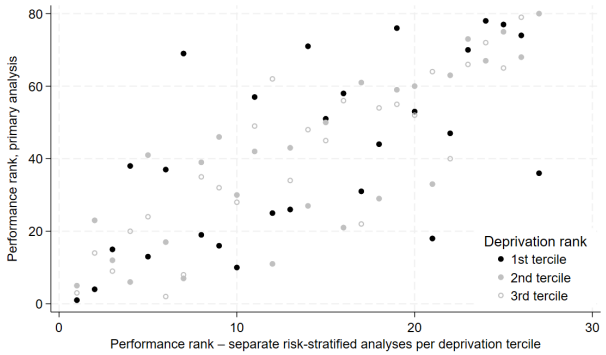

Figure 7 PCP ranks from the risk-stratified analyses against ranks from the primary specification

to systematic differences between PCPs: the ICC increases to 56.0% (95% CI: [41.7%; 69.5%]) from 51.9% (95% CI: [37.8%; 65.7%]). This increase implies that approximately 4.1 percentage points of the between-PCP variation captured in the alternative specification reflect socioeconomic differences that are accounted for by the deprivation index in the baseline model.

Using the performance metric $\widehat{u}_g$ from this alternative specification, we compute PCP performance rankings and compare them with those from the primary analysis. The resulting rank correlation is high (0.915), indicating that the identification of relatively high- and low-performing practices is largely unaffected. This evidence supports the robustness of the model and suggests that the main findings are not driven by unobserved correlations between deprivation and PCP

performance.

6. Robustness checks

In this section we test the robustness of several sample-defining choices.

6.1. Alternative thresholds to define relevant PCPs

The primary analysis focuses on the top PCPs accounting for 50% of attendances at Hospital 1 and 65% at Hospital 2, resulting in 81 distinct PCPs. In this section we test the robustness of our findings with respect to these thresholds. More specifically, we re-estimate models (1) – (4) for

- the top PCPs accounting for 55% of attendances at Hospital 1 and 60% at Hospital 2, resulting in 80 distinct PCPs.
- the top PCPs accounting for 55% of attendances at Hospital 1 and 70% at Hospital 2, resulting in 91 distinct PCPs.
- the top PCPs accounting for 60% of attendances at Hospital 1 and 60% at Hospital 2, resulting in 90 distinct PCPs.
- the top PCPs accounting for 60% of attendances at Hospital 1 and 70% at Hospital 2, resulting in 101 distinct PCPs.

All alternative samples yield similar results to those presented in the main paper (see Tables 12 – 17). We therefore conclude that the findings are not driven by the discretionary thresholds.

6.2. Excluding PCP closures and expanding PCPs

In the primary analysis we include data from PCPs that expanded to new branches up until one year before the expansion. To alleviate concerns that the results may be driven by expansions, we now exclude expanding (N=1) PCPs from the entire study period and re-estimate models (1)–(4). The results are presented in Tables 12 – 17 and are quantitatively comparable to the primary analysis presented in the paper.

Table 12 Estimating model (1) on different samples

	Top 80 PCPs	Top 91 PCPs	Top 90 PCPs	Top 101 PCPs	Excluding expanding PCPs
	(1)	(2)	(3)	(4)	(5)
	A_{gkt}	A_{gkt}	A_{gkt}	A_{gkt}	A_{gkt}
Closest hospital	0.003	0.004	0.002	0.003	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Deprivation rank	-0.001**	-0.001**	-0.001***	-0.001**	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Scale	-0.000	-0.000	-0.000	-0.000	0.000
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)
Female	0.125	0.122	0.128	0.119	0.138
	(0.108)	(0.105)	(0.113)	(0.111)	(0.100)
Elderly	0.011	-0.007	-0.111	-0.098	-0.166
	(0.174)	(0.136)	(0.145)	(0.131)	(0.162)
CMI	0.025	0.045	0.030	0.044	0.060*
	(0.039)	(0.033)	(0.040)	(0.035)	(0.033)
Constant	Yes	Yes	Yes	Yes	Yes
ED × Year FE	Yes	Yes	Yes	Yes	Yes
$\widehat{\tau}^2$	0.0002358	0.0002041	0.0002334	0.0002192	0.0001736
$\widehat{\tau}^2$ 95% CI	[.0001445; .0003846]	[.0001353; .0003079]	[.0001519; .0003587]	[.0001497; .0003209]	[.0001046; .0002878]
ICC	56.6%	51.5%	49.2%	47.9%	51.80%
ICC 95% CI	[41.8%; 70.3%]	[39.7%; 63.2%]	[36.0%; 62.5%]	[36.7%; 59.4%]	[37.7%; 65.7%]
Wald χ^2	502.9	743.0	560.4	881.2	918.6
Prop $> \chi^2$	0.000	0.000	0.000	0.000	0.000
Log Pseudolikelihood	1,154.8	1,422.3	1,282.7	1,551.2	1,163.3
Observations	428	528	497	597	420
Number of groups	80	91	90	101	79

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

Table 13 Estimating model (2) on different samples

	Top 80 PCPs	Top 91 PCPs	Top 90 PCPs	Top 101 PCPs	Excluding expanding PCPs
	(1)	(2)	(3)	(4)	(5)
	Not rec	Not rec	Not rec	Not rec	Not rec
$\widehat{u}_g$	1.021**	0.991**	1.119***	1.036***	1.241***
	(0.406)	(0.398)	(0.358)	(0.339)	(0.463)
$\widehat{\epsilon}_{gt}$	0.084	0.036	0.088	0.045	-0.008
	(0.203)	(0.189)	(0.160)	(0.153)	(0.216)
Closest hospital	-0.004	-0.008	-0.005	-0.009	-0.009
	(0.016)	(0.016)	(0.015)	(0.015)	(0.017)
Deprivation rank	0.000	0.000	-0.000	-0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Scale	-0.003**	-0.003**	-0.003**	-0.003**	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Female	-0.073	-0.094	-0.072	-0.095	-0.147
	(0.119)	(0.113)	(0.110)	(0.106)	(0.123)
Elderly	1.291**	1.124**	0.664	0.661*	1.659***
	(0.559)	(0.462)	(0.423)	(0.363)	(0.503)
CMI	-0.196	-0.189*	-0.147	-0.147	-0.242**
	(0.120)	(0.107)	(0.113)	(0.102)	(0.107)
Constant	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
PCP effect	Random	Random	Random	Random	Random
Wald χ^2	24.7	24.9	23.4	24.0	21.3
Prop $> \chi^2$	0.016	0.015	0.023	0.020	0.046
Log Pseudolikelihood	598.2	687.9	675.9	766.4	610.3
Observations	384	439	433	488	384
Number of groups	80	91	90	101	79

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
Not rec: Proportion of patients refraining to recommend their PCP.

Table 14 Estimating model (2) on different samples

	Top 80 PCPs	Top 91 PCPs	Top 90 PCPs	Top 101 PCPs	Excluding expanding PCPs
	(1)	(2)	(3)	(4)	(5)
	No access	No access	No access	No access	No access
$\widehat{u}_g$	0.955***	0.889***	1.369***	1.238***	1.040***
	(0.336)	(0.343)	(0.323)	(0.323)	(0.369)
$\widehat{\epsilon}_{gt}$	0.508**	0.362*	0.349**	0.248	0.520**
	(0.211)	(0.202)	(0.165)	(0.163)	(0.240)
Closest hospital	0.002	-0.005	0.002	-0.003	-0.006
	(0.014)	(0.014)	(0.013)	(0.013)	(0.015)
Deprivation rank	-0.003**	-0.002**	-0.002**	-0.002**	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Scale	-0.001	-0.001	0.000	-0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Female	0.318***	0.300***	0.224**	0.217**	0.305***
	(0.106)	(0.103)	(0.112)	(0.109)	(0.115)
Elderly	1.272***	0.973**	0.827**	0.673**	1.328***
	(0.449)	(0.406)	(0.331)	(0.306)	(0.452)
CMI	-0.308***	-0.258***	-0.228**	-0.194**	-0.301***
	(0.090)	(0.083)	(0.091)	(0.083)	(0.088)
Constant	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
PCP effect	Random	Random	Random	Random	Random
Wald χ^2	64.5	57.2	61.0	54.7	64.0
Prop $> \chi^2$	0.000	0.000	0.000	0.000	0.000
Log Pseudolikelihood	605.2	682.8	680.5	757.9	602.5
Observations	384	439	433	488	384
Number of groups	80	91	90	101	79

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
No access: Proportion of patients experiencing access problems at their PCP.

Table 15 Estimating model (3) on different samples

	Top 80 PCPs	Top 91 PCPs	Top 90 PCPs	Top 101 PCPs	Excluding expanding PCPs
	(1) $P(QOF_{gt} \geq 90\%)$	(2) $P(QOF_{gt} \geq 90\%)$	(3) $P(QOF_{gt} \geq 90\%)$	(4) $P(QOF_{gt} \geq 90\%)$	(5) $P(QOF_{gt} \geq 90\%)$
$\widehat{u}_g$	-18.184	-19.552	-6.053	-2.895	-39.599**
	(12.681)	(13.570)	(15.470)	(15.649)	(15.693)
$\widehat{\epsilon}_{gt}$	16.933	14.448	5.300	3.797	12.967
	(11.160)	(10.681)	(8.229)	(8.087)	(11.467)
Closest hospital	0.880	0.894	0.944	1.007	0.780
	(0.716)	(0.706)	(0.726)	(0.716)	(0.696)
Deprivation rank	-0.009	-0.021	-0.009	-0.021	-0.026
	(0.040)	(0.038)	(0.035)	(0.033)	(0.038)
Scale	0.010	-0.026	0.050	0.019	-0.023
	(0.060)	(0.057)	(0.062)	(0.058)	(0.055)
Female	3.961	4.513	3.854	4.336	6.576*
	(3.920)	(3.827)	(3.480)	(3.287)	(3.819)
Elderly	-13.933	-10.853	-25.477*	-20.872	-25.717
	(18.693)	(16.766)	(14.726)	(14.490)	(19.241)
CMI	-1.854	-2.457	-1.669	-2.405	-0.819
	(4.012)	(3.710)	(3.722)	(3.500)	(3.864)
Constant	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
PCP effect	Random	Random	Random	Random	Random
Wald χ^2	21.8	23.1	25.3	26.4	26.4
Prop $> \chi^2$	0.040	0.027	0.013	0.009	0.009
Log Pseudolikelihood	-113.4	-128.0	-134.8	-150.0	-114.4
Observations	385	440	434	489	386
Number of groups	80	91	90	101	79

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

Table 16 Estimating model (3) on different samples

	Top 80 PCPs	Top 91 PCPs	Top 90 PCPs	Top 101 PCPs	Excluding expanding PCPs
	(1)	(2)	(3)	(4)	(5)
	$P(QOF_{gt} \geq 95\%)$	$P(QOF_{gt} \geq 95\%)$	$P(QOF_{gt} \geq 95\%)$	$P(QOF_{gt} \geq 95\%)$	$P(QOF_{gt} \geq 95\%)$
$\widehat{u}_g$	-8.701	-8.431	-3.980	-0.678	-10.643
	(12.758)	(13.453)	(13.821)	(13.742)	(17.380)
$\widehat{\epsilon}_{gt}$	19.986**	16.067**	12.731*	9.768	19.067**
	(8.297)	(7.880)	(7.080)	(6.791)	(8.722)
Closest hospital	0.668	0.724	0.698	0.778	0.640
	(0.548)	(0.534)	(0.545)	(0.531)	(0.555)
Deprivation rank	-0.009	-0.020	-0.013	-0.023	-0.013
	(0.036)	(0.035)	(0.034)	(0.033)	(0.042)
Scale	-0.016	-0.037	-0.001	-0.018	-0.044
	(0.062)	(0.059)	(0.058)	(0.056)	(0.066)
Female	5.530	6.505	6.491	7.395	7.558
	(5.843)	(5.526)	(5.438)	(5.083)	(5.737)
Elderly	10.506	15.850	0.625	6.635	5.099
	(16.197)	(15.205)	(13.401)	(13.540)	(18.118)
CMI	-4.085	-5.703	-4.801	-6.251*	-4.505
	(3.806)	(3.560)	(3.754)	(3.542)	(3.826)
Constant	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
PCP effect	Random	Random	Random	Random	Random
Wald χ^2	28.3	31.2	21.0	25.0	25.3
Prop $> \chi^2$	0.005	0.002	0.050	0.015	0.014
Log Pseudolikelihood	-204.5	-227.5	-228.4	-251.2	-201.1
Observations	385	440	434	489	386
Number of groups	80	91	90	101	79

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

Table 17 Estimating model (4) on different samples

	Top 80 PCPs	Top 91 PCPs	Top 90 PCPs	Top 101 PCPs	Excluding expanding PCPs
	(1)	(2)	(3)	(4)	(5)
	$\widehat{u}_g$	$\widehat{u}_g$	$\widehat{u}_g$	$\widehat{u}_g$	$\widehat{u}_g$
PpFTE	0.009***	0.008***	0.008***	0.007***	0.008***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Closest hospital	-0.006*	-0.005*	-0.006*	-0.005	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Deprivation rank	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Scale	0.000	-0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	0.003	0.006	-0.021	-0.009	0.027
	(0.071)	(0.068)	(0.083)	(0.080)	(0.048)
Elderly	-0.194	-0.099	-0.166	-0.127	-0.174
	(0.157)	(0.143)	(0.106)	(0.113)	(0.137)
CMI	0.019	0.014	0.039	0.032	0.012
	(0.037)	(0.030)	(0.034)	(0.030)	(0.030)
Constant	Yes	Yes	Yes	Yes	Yes
Observations	79	90	88	99	78
R^2	0.217	0.175	0.185	0.145	0.249

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
As in the main paper, we excluded PCPs with implausible staffing numbers (one PCP excluded in (1), (2) and (5), two PCPs excluded in (3)-(4)).

7. Scalability of the Approach: From Single Hospitals to National Data

The approach described in §3 of the paper 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.

7.1. Consistency from One to Two Hospitals

To demonstrate the approach'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 of the paper, we estimate model (1) independently for Hospital 1 and Hospital 2. Table 18 presents the results and Figure 8 shows the $\hat{u}_g$ distribution for both hospitals.

The results indicate greater uncertainty in estimating between-PCP variability when analyzing each hospital independently compared to the combined analysis, leading to wider 95% CIs. Specifically, the ICC estimated based on Hospital 1 is 52.6% (95% CI [29.2%; 75.1%]), and for Hospital 2 is 33.9% (95% CI [22.6%; 47.6%]). Crucially, despite this increased uncertainty, the individual estimates of PCP burden on EDs ($\hat{u}_g$) are highly consistent between the single-hospital and combined analyses. Figure 9 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 approach is likely to remain robust when scaled up to a larger number of hospitals, potentially at a national level.

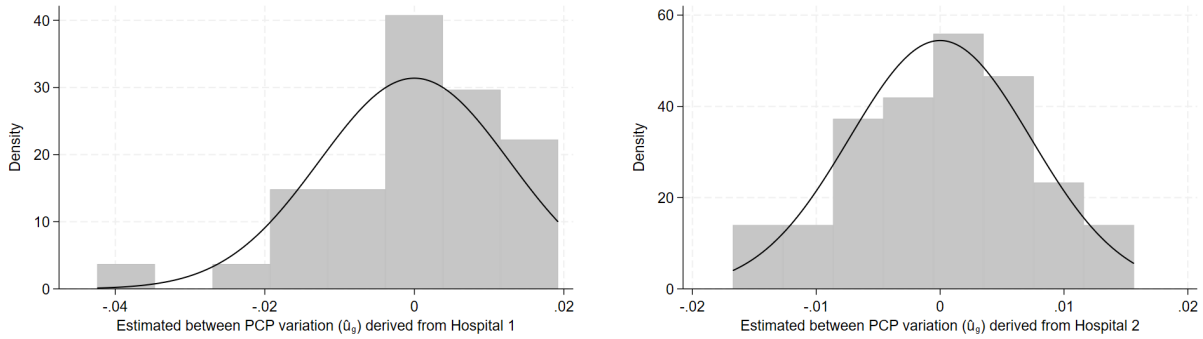

Figure 8 Variation between PCPs.

Subsequently, we replicate models (2) – (4) to demonstrate that the measure $\widehat{u}_g$ estimated using single hospital data is correlated with PCP surveys, QOF scores, and patient-to-staff ratios. Not surprisingly, since the single hospital analysis uses less data, some of the results are estimated with more uncertainty compared to the analysis using both hospitals. Specifically, we find that PCPs

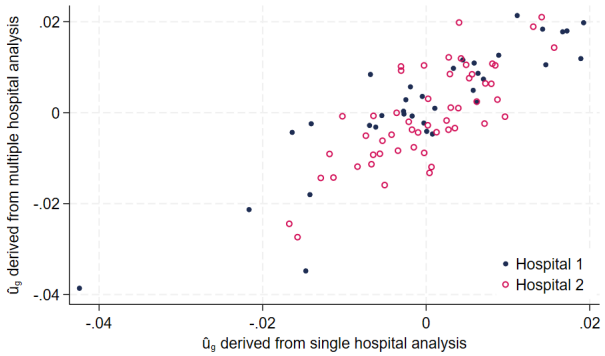

Figure 9 Correlation between the U_g

with a higher $\widehat{u}_g$ are also the ones patients are more likely dissatisfied with. Table 19 indicates $\beta_U > 0$ for all models, albeit in Column (1) not distinguishable from zero. Regarding the QOF score, most of the coefficients γ_U are negative as expected but they are estimated with very large standard errors leading to results that are not statistically significant (Table 20). For the performance antecedents we find, as expected, a positive association between patient-to-staffing ratio and $\widehat{u}_g$ for both hospitals (Table 21). Taken together the results provide similar indications of validity as the analysis conducted on the joint data.

Table 18 Estimating model (1) for both hospitals separately

	Hospital 1 (1)	Hospital 2 (2)
Closest hospital	-0.016***	0.000
	(0.006)	(0.006)
Deprivation rank	-0.002**	-0.001**
	(0.001)	(0.000)
Scale	0.001	0.000
	(0.001)	(0.000)
Female	0.172***	-0.213***
	(0.063)	(0.066)
Elderly	-0.151	-0.406***
	(0.351)	(0.150)
CMI	0.143*	0.065**
	(0.082)	(0.025)
Constant	Yes	Yes
Year FE	Yes	Yes
$\widehat{\tau}^2$	0.0001875	0.0000736
$\widehat{\tau}^2$ 95% CI	[.0000782; .0004494]	[.0000441; .000123]
ICC	52.7%	33.9%
ICC 95% CI	[29.2%; 75.1%]	[22.6%; 47.6%]
Wald χ^2	136.1	179.4
Prop $> \chi^2$	0.000	0.000
Log Pseudolikelihood	450.8	751.3
Observations	165	261
Number of groups	35	53

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

Table 19 Estimating model (2) for both hospitals separately

	Hospital 1		Hospital 2	
	(1)	(2)	(3)	(4)
	Not rec	No access	Not rec	No access
$\widehat{u}_g$	1.024	0.921**	1.915**	2.037***
	(0.670)	(0.449)	(0.772)	(0.727)
$\widehat{\epsilon}_{gt}$	0.142	0.776**	-0.096	0.197
	(0.355)	(0.316)	(0.196)	(0.281)
Closest hospital	0.020	0.022	-0.035	-0.025
	(0.019)	(0.016)	(0.021)	(0.019)
Deprivation rank	-0.001	-0.004**	-0.000	-0.003***
	(0.002)	(0.002)	(0.001)	(0.001)
Scale	0.001	0.003**	-0.003**	-0.002
	(0.002)	(0.001)	(0.001)	(0.001)
Female	-0.183	0.299**	-0.640**	-0.254
	(0.115)	(0.124)	(0.309)	(0.260)
Elderly	2.902***	2.561***	0.934*	0.992*
	(1.106)	(0.699)	(0.482)	(0.598)
CMI	-0.449**	-0.620***	-0.271**	-0.277***
	(0.212)	(0.147)	(0.125)	(0.105)
Constant	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
PCP effect	Random	Random	Random	Random
Wald χ^2	17.9	113.4	37.5	30.6
Prop $>\chi^2$	0.118	0.000	0.000	0.002
Log Pseudolikelihood	266.6	257.0	418.0	421.1
Observations	162	162	261	261
Number of groups	35	35	53	53

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

7.2. Adapting the Approach for National-Level Data

Scaling the approach to national-level data, encompassing all hospitals and PCPs, requires some adaptations to the approach outlined in the paper. 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.

7.2.1. 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 of the paper, 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 §6.1, it would remain important to assess the robustness of the results to different threshold values.

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

7.2.2. Hospital Fixed or Random Effects? In the primary analysis outlined in the main paper, we used hospital ED fixed effects (ED_k dummy variables) to control for structural differences

Table 20 Estimating model (3) for both hospitals separately

	Hospital 1		Hospital 2	
	(1)	(2)	(3)	(4)
	$P(QOF_{gt} \geq 90\%)$	$P(QOF_{gt} \geq 95\%)$	$P(QOF_{gt} \geq 90\%)$	$P(QOF_{gt} \geq 95\%)$
$\widehat{u}_g$	-30.024	9.945	-99.535***	-23.309
	(23.629)	(21.886)	(30.325)	(37.208)
$\widehat{\epsilon}_{gt}$	8.994	25.058**	15.249	4.214
	(17.450)	(10.381)	(14.449)	(11.900)
Closest hospital	0.638	1.446*	0.661	-0.516
	(0.922)	(0.773)	(1.034)	(0.913)
Deprivation rank	0.034	-0.060	-0.048	-0.027
	(0.077)	(0.064)	(0.045)	(0.060)
Scale	-0.005	0.104	-0.028	-0.173*
	(0.082)	(0.070)	(0.061)	(0.091)
Female	6.712*	2.201	1.645	7.847
	(3.863)	(5.475)	(13.295)	(15.283)
Elderly	-44.601	10.765	-20.077	-3.489
	(29.225)	(26.732)	(21.060)	(25.136)
CMI	5.439	-5.087	-0.158	-5.056
	(7.284)	(6.196)	(4.424)	(5.065)
Constant	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
PCP effect	Random	Random	Random	Random
Wald χ^2	39.5	26.8	29.3	20.7
Prop $> \chi^2$	0.000	0.005	0.004	0.055
Log Pseudolikelihood	-35.9	-83.2	-81.4	-129.6
Observations	135	164	261	261
Number of groups	35	35	53	53

PCP-Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
The number of observations differ between (1) and (2) due to perfect prediction.

Table 21 Estimating model (4) for both hospitals separately

	Hospital 1	Hospital 2
	(1)	(2)
	$\widehat{u}_g$	$\widehat{u}_g$
PpFTE	0.010***	0.005***
	(0.003)	(0.001)
Closest hospital	-0.005	0.002
	(0.005)	(0.004)
Deprivation rank	0.000	-0.000
	(0.001)	(0.000)
Scale	-0.000	0.000
	(0.000)	(0.000)
Female	0.016	0.123*
	(0.039)	(0.072)
Elderly	-0.293	-0.027
	(0.317)	(0.112)
CMI	0.040	0.002
	(0.076)	(0.020)
Constant	Yes	Yes
Observations	35	52
R^2	0.291	0.242

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.
As in the main paper, we excluded the one PCP with implausible staffing numbers, leading to N=52 for Hospital 2.

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.

7.2.3. 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 of the main paper), 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.

7.3. Implementation Considerations

Scaling the proposed approach 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 approach'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 the paper demonstrates the approach'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.

8. Bootstrapping the effect size

In §6 of the main paper we estimate the impact on ACS attendances and costs of a counterfactual scenario where PCPs that performed worse than the 25% could improve their performance to the 25% percentile. These results were estimated using the following bootstrapping algorithm:

1. We draw a random sample of 80 PCPs with replacement, ensuring that the resampling respects the natural grouping of the data, specifically the proportion of ACS attendances A_{gkt} within each PCP g. We chose 80 PCPs because this matches the number of PCP clusters used in the primary analysis. Since PCP clusters are sampled with replacement, some clusters may appear more than once, while others may not appear at all in a given bootstrap sample. As a result, the number of observations in the bootstrap sample may differ from the original, depending on how many times individual clusters are selected and their size.

2. We estimate model (1) based on this new dataset.

3. We assume that for each PCP-hospital-year the number of Non-ACS attendances remains as observed, and calculate the predicted number of ACS attendances for each PCP-year, using $ACS = \frac{\widehat{A}_{gkt}}{1-\widehat{A}_{gkt}} NonACS$.

4. We aggregate these predictions to the ED level to obtain the status quo predictions.

5. We determine the 25% percentile of $\widehat{u}_g$, denoted as $\widehat{u}_{25}$.

6. For PCPs that perform worse than the 25% percentile (i.e. $\widehat{u}_g > \widehat{u}_{25}$), we set $\widehat{u}_g$ equal to $\widehat{u}_{25}$. PCPs that perform better than the 25% percentile are left unchanged.

7. We repeat steps 3 to 4 to obtain the predicted number of ACS attendances for the counterfactual scenario.

8. We calculate the difference in ACS attendances compared to the status quo.

9. We repeat steps 1 to 8 10,000 times to estimate a 95% confidence interval for the difference in ACS attendances based on the empirical distribution.

9. Determining the impact of different resource allocation policies on ACS ED attendances

In this section we estimate the impact of four different resource allocation policies on ACS ED attendances. Each allocation policy uses a different criterion to identify PCPs that are selected to receive an additional 0.5 FTE physicians:

- Policy 1 allocates resources to PCPs with the highest proportion of ACS attendances (as measured by $\widehat{u}_g$).
- Policy 2 selects PCPs with the highest proportion of patients reporting access problems (averaged over the duration of study).
- Policy 3 allocates resources according to the PCP's proportion of ACS admissions (as measured according to §4.4 of the paper).
- Policy 4 is a random policy in which PCPs are selected randomly with equal probability.

For each policy the results were estimated according to the steps below. Note that for the random policy 4 we repeat the steps 1,000 times to derive the mean of the counterfactual predictions.

1. We calculate the impact at PCP-level averaged across EDs and years: First, we use the model estimates from (1) to compute the PCP's adjusted proportion of ACS visits: $\widehat{A}_{gkt} = \widehat{\alpha}_0 + \widehat{\alpha}_C C_{gkt}$, average this over the duration of study, and denote this as $\tilde{A}_g$. $\tilde{A}_g$ is the averaged part of PCP performance that is not related to staffing and therefore not affected by the allocation policies.

2. We assume that $X = \{1, \ldots, 15\}$ FTE can be allocated in increments of 0.5 FTE across different PCPs. This means that $N = X/0.5$ PCPs will see an increase in staffing.

3. We rank PCPs in descending order of the selection criterion and select the first N PCPs to see an increase in staffing.

4. We compute counterfactual patient-to-staff ratios $PpFTE_g^c$. Specifically, for PCPs that are ranked $\leq N$, we set their staffing $PpFTE_g^c$ equal to the average number of patients per physician FTE and add 0.5 FTE to the denominator. For PCPs that are ranked $> N$, their staffing levels are left unchanged, i.e. $PpFTE_g^c = PpFTE_g$.

5. Subsequently, we use the model estimates from (4) to compute the counterfactual random effect $\tilde{u}_g = \widehat{\delta}_0 + \widehat{\delta}_P PpFTE_g^c + \widehat{\delta}_C C_g$.

6. We assume that for each PCP the average number of Non-ACS attendances remains as observed, and calculate the predicted number of ACS attendances for each PCP, using $ACS = \frac{(\tilde{A}_g + \tilde{u}_g)}{1-(\tilde{A}_g + \tilde{u}_g)} NonACS$.

7. We aggregate these predictions to the ED level to obtain the counterfactual predictions at annual ED level.

The results of this exercise are presented in Figure 10. Figure 10a illustrates the number of annual ACS attendances per ED for different allocation policies as the number of additional FTEs increases. Figure 10b shows the relative performance of each policy compared to random allocation. The findings suggest that allocating resources based on the measure ($\widehat{u}_g$) could lead to a more significant reduction in ACS attendances compared to other allocation strategies. Notably, the measure ($\widehat{u}_g$) 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.

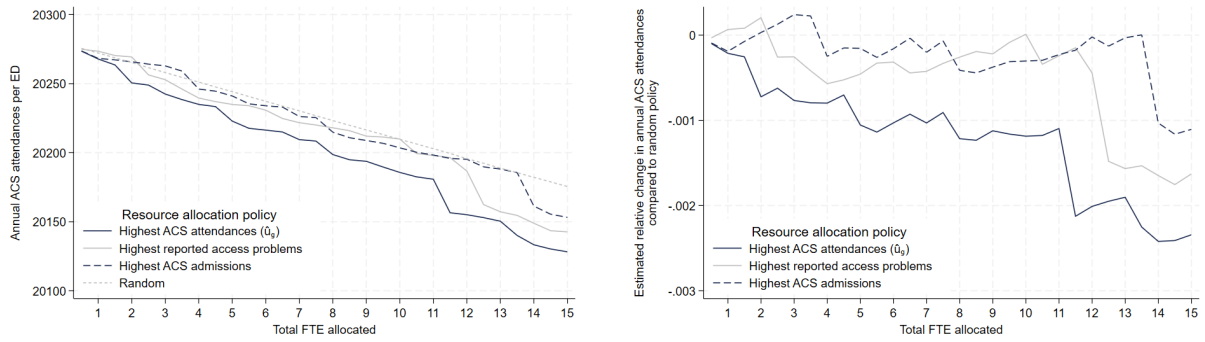

Figure 10 Impact of different resource allocation policies on ACS ED attendances.

References

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

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.