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

(Authors' names blinded for peer review)

This online appendix includes additional information and robustness checks on the analysis performed in the main paper.

Contents

1. Model assumptions primary analysis

The model presented in (1) assumes $u_g \sim N(0, \tau^2)$ and $\epsilon_{gt} \sim N(0, \sigma^2)$. To empirically assess the validity of these assumptions, we examine the residuals $\widehat{u}_g$ and $\widehat{\epsilon}_{gt}$ using the results of the full model reported in Table 5, column 3. Figure 1 shows the empirical distribution plotted against the theoretical normal distribution (qq-plot). The figures show that the normal distribution is a good fit. We additionally conduct the skewness-kurtosis test for normality of the random effect u_g and idiosyncratic error ϵ_{gt}. The test does not find sufficient evidence to reject the normality assumption for u_g ($p = 0.2531$). For ϵ_{gt} the normality assumption is rejected (p<0.001). Therefore we calculate robust standard errors clustered at the PCP level which allows for errors to be heteroskedastic across PCP and correlated within PCP.

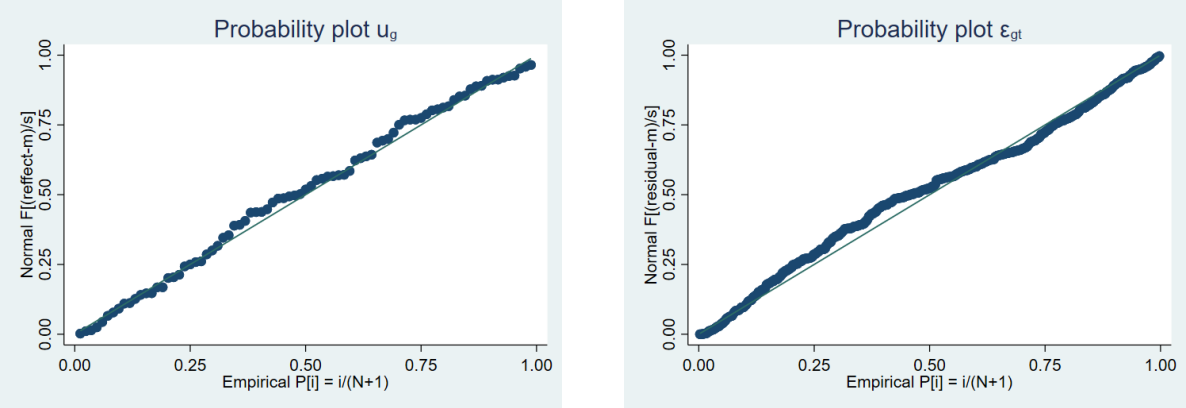


Figure 1 Theoretical distribution vs. empirical distribution

1.1. Mundlak test for orthogonality between u_g and explanatory variables

The random-effect model of (1) assumes that the unobserved systematic differences between PCPs u_g are orthogonal to the explanatory variables. This is an assumption that could be violated if, for example larger PCPs are more likely to invest in resources that allow them to be more effective in treating patients. In this case, a fixed-effect model would be more appropriate (Wooldridge 2010). To test whether there is evidence that this assumption is violated we perform a test based on (Mundlak 1978). This is an alternative to the more widely used Hausman test for random vs fixed effect. In contrast to the Hausman test, this test allows for heteroskedastic errors within PCP and intragroup correlation. The test involves adding the mean of the time-varying variables of equation (1) (i.e., scale, proportions of female and elderly, CMI; see Table 1 column (4) for the results of this model specifcation) as explanatory variables and testing if their coefficients are jointly different to zero. Failing to reject this hypothesis constitutes evidence for the random- and against the fixed-effect specification. The p-value of this test is 13.4% which suggests that the random-effect specification is appropriate.

2. Alternative model specifications

For the model presented in (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 heterogeneity that might affect ACS patients differently than Non-ACS patients. In this section we investigate three alternative modeling choices.

2.1. 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 $Scale_{gt}^2$ to the model. For comparison, Column (1) in Table 1 repeats the results of the main paper and Column (2) shows the results of the model with $Scale_{gt}^2$. There is no evidence of non-linear scale effects as the coefficients of $Scale_{gt}$ and $Scale_{gt}^2$ are not statistically significant.

2.2. Testing the logarithmic specification

The dependent variable is the proportion of ACS attendances $A_{gt} = \frac{ACS_{gt}}{ACS_{gt}+NonACS_{gt}}$ and we log-transformed this variable for estimation:

$$ln(A_{gt}) = \alpha_0 + \alpha_C C_{gt} + u_g + \epsilon_{gt}. \tag{1}$$

In this section we test whether one aspect of this specification is correct. More specifically, since

$$ln(A_{gt}) = ln(\frac{ACS_{gt}}{ACS_{gt}+NonACS_{gt}}) = ln(ACS_{gt}) - ln(ACS_{gt}+NonACS_{gt}) \tag{2}$$

we can substitute (2) in (1) to yield

$$ln(ACS_{gt}) = \beta_0 + \beta_1 ln(ACS_{gt}+NonACS_{gt}) + \beta_C C_{gt} + u_g + \epsilon_{gt}. \tag{3}$$

If the model specification of (1) is correct, the estimated coefficient $\widehat{\beta}_1$ in model (3) should be close to 1. Column (3) in Table 1 presents the results of estimating model (3). Indeed, $\widehat{\beta}_1 = 1.076$ (standard error = 0.012) and the rest of the results are similar to the primary model specification, column (1).

2.3. 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

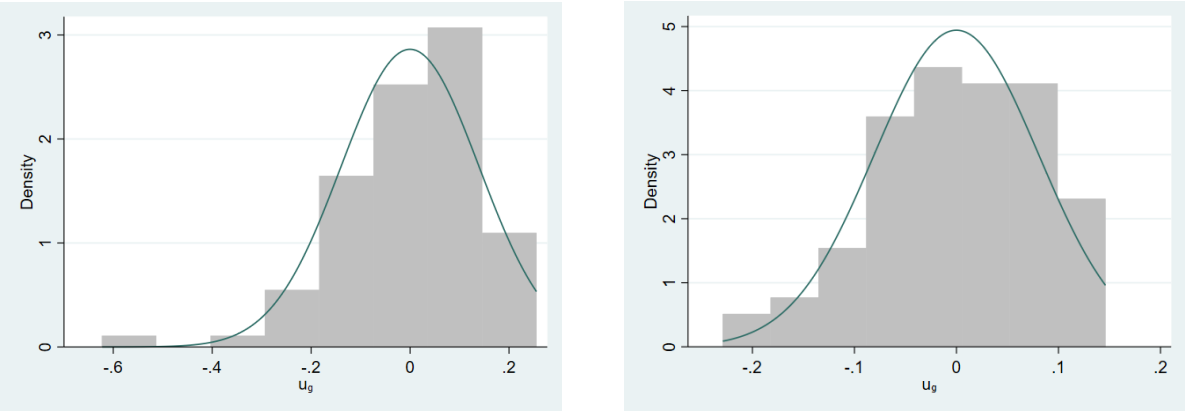
Table 1 Decomposing variation in PCP performance: Alternative model specifications

	(1) $\ln(A_{gt})$	(2) $\ln(A_{gt})$	(3) $\ln(ACS_{gt})$	(4) $\ln(A_{gt})$	(5) $\ln(A_{gt})$
Closest hospital	0.024	0.030	−0.059∗	0.010	
	(0.022)	(0.022)	(0.029)	(0.025)	
Depriviation rank	−0.007∗	−0.007∗	−0.005∗	−0.005	
	(0.003)	(0.003)	(0.003)	(0.003)	
Scale	−0.004	−0.019	−0.010∗∗	0.006	0.007
	(0.004)	(0.013)	(0.003)	(0.007)	(0.007)
Scale2		0.001			
		(0.001)			
Female	0.846	0.977	1.245∗	−0.124	−0.119
	(0.657)	(0.666)	(0.584)	(0.761)	(0.768)
Elderly	−0.151	−0.281	−0.362	0.950	1.042
	(0.789)	(0.803)	(0.691)	(2.786)	(2.802)
CMI	0.011	−0.001	0.060	−0.383	−0.386
	(0.231)	(0.216)	(0.244)	(0.299)	(0.301)
Ln(#Non-Acs+Acs)			1.076∗∗∗		
			(0.012)		
Mean Scale				−0.011	
				(0.008)	
Mean Female				1.026	
				(0.951)	
Mean Elderly				−1.827	
				(3.002)	
Mean CMI				0.709	
				(0.437)	
Constant	−1.958∗∗∗	−1.942∗∗∗	−2.637∗∗∗	−2.256∗∗∗	−1.277∗∗∗
	(0.354)	(0.350)	(0.358)	(0.388)	(0.322)
Year FE	Yes	Yes	Yes	Yes	Yes
PCP effect	Random	Random	Random	Random	Fixed
$\widehat{\tau}^2$ 95% CI	[0.006; 0.013]	[0.006; 0.013]	[0.003; 0.010]	[0.006; 0.012]	na
Intraclass correlation (ICC)	36.96%	36.73%	26.99%	36.63%	56.20%
Model Wald χ^2	45.60	52.73	14,124.48	62.19	na
Observations	401	401	401	401	401
Number of groups	83	83	83	83	83

Clustered standard errors in parentheses. na: statistics not provided in the fixed-effects estimation.
*** p<0.001, ** p<0.01, * p<0.05.

is orthogonal to any of the time varying controls included in the model. We note that this model specification cannot include the distance variable $closest_g$ 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 (1) column (5). In the fixed-effect specification 56.20% of the variation is attributed to systematic differences between PCPs compared to the 36.96% in the random-effect specification (Table (1) 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



(a) Fixed-effect specification (b) Random-effect specification

Figure 2 Variation between PCPs

than the random-effect specification. This is illustrated in Figure 2. In fact, the variance of $\widehat{u}_g$ is 0.019 in the fixed-effect specification is more than twice as high as the variance (0.009) of the random-effect specification.

Nevertheless, the PCP-effect estimated by the fixed effect model is highly correlated to that estimated by the random effect model (correlation=0.817, p<0.001) which indicates that the two specifications overlap in identifying PCP that perform better (or worse) than average. Furthermore, the effect calculated using the fixed effect model is very much correlated with patient surveys. More specifically, the results of estimating model (5) with $\widehat{u}_g$ obtained through the fixed-effect specification appear in Table 2 and are similar the the results reported in the main paper. Hence, using a fixed-effect specification does not alter the paper's conclusions.

3. Robustness Checks

In this section we test the robustness of several sample-defining choices. We re-estimate the performance decomposition model of §3:

$$ln(A_{gt}) = \alpha_0 + \alpha_C C_{gt} + u_g + \epsilon_{gt}, \tag{4}$$

and for validation purposes, we re-estimate the model using the results of the patient survey presented in §4.1:

$$P^i_{gt} = \beta_0 + \beta_U \widehat{u}_g + \beta_\epsilon \widehat{\epsilon}_{gt} + \beta_C C_{gt} + \nu_g + \epsilon^P_{gt}, \tag{5}$$

3.1. Alternative thresholds to define TOP-PCP samples

In the primary analysis we focussed on the subsample of PCPs that collectively account for 75% of the ED's ACS attendances. In this section we test the robustness of our findings with respect to

Table 2 Estimating model (5) (validation) using $\widehat{u}_g$ obtained from a fixed-effect specification

	(1) NotRec	(2) NoAccess
$\widehat{u}_g$	0.080	0.168***
	(0.050)	(0.040)
$\widehat{\epsilon}_{gt}$	0.012	−0.012
	(0.015)	(0.020)
Closest hospital	−0.008	−0.003
	(0.020)	(0.016)
Depriviation rank	0.003	0.000
	(0.001)	(0.001)
Scale	−0.002	0.003*
	(0.002)	(0.001)
Female	−0.016	−0.055
	(0.155)	(0.118)
Elderly	−0.075	0.402
	(0.360)	(0.296)
CMI	0.064	−0.148*
	(0.086)	(0.062)
Constant	0.018	0.262***
	(0.120)	(0.075)
Year FE	Yes	Yes
Variance ν_g^P 95% CI	[0.001; 0.005]	[0.001; 0.002]
Intraclass correlation (ICC)	62.73%	47.39%
Model Wald χ^2	18.55	44.80
Observations	397	397
Number of groups	83	83

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
NotRec: Proportion of patients refraining to recommend their PCP. NoAccess: Proportion of patients experiencing access problems at their PCP

this threshold. More specifically, we re-estimate model (4) and re-assess validity through estimating model (5) for i) the top 136 PCPs that collectively account for 80% for the ED's ACS attendances; ii) the top 60 PCPs that together account for 70% of ACS attendances. Both alternative samples yield similar results as presented in the main paper (see Tables 3 and 4 for the case of 136 PCPs and Tables 5 and 6 for the case of 60 PCPs, respectively). We therefore conclude that our findings are not driven by the discretionary threshold of 75% ACS ED attendances.

Table 3 Estimating model (4) (variance decomposition) using data from the top 136 PCPs that collectively account for 80% of ACS attendances

	(1) $\ln(A_{gt})$	(2) $\ln(A_{gt})$	(3) $\ln(A_{gt})$
Closest hospital		0.097***	0.099***
		(0.026)	(0.029)
Depriviation rank		−0.012***	−0.009**
		(0.003)	(0.003)
Scale			−0.014*
			(0.006)
Female			0.359
			(0.840)
Elderly			−2.026*
			(1.016)
CMI			0.181
			(0.250)
Constant	−1.725***	−1.601***	−1.807***
	(0.023)	(0.040)	(0.435)
Year FE	Yes	Yes	Yes
$\widehat{\tau}^2$ 95% CI	[0.026; 0.049]	[0.020; 0.042]	[0.018; 0.034]
Intraclass correlation (ICC)	51.57%	46.24%	41.64%
Model Wald χ^2	19.96	81.96	92.18
Observations	666	652	645
Number of groups	136	133	132

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
Note: Number of observation differs due to missing location and case-mix information in 19 PCP-years.

Table 4 Estimating model (5) (validation) using data from the top 136 PCPs that collectively account for 80% of ACS attendances

	(1) NotRec	(2) NoAccess
$\widehat{u}_g$	0.055	0.069∗
	(0.030)	(0.028)
$\widehat{\epsilon}_{gt}$	0.003	−0.013
	(0.008)	(0.011)
Closest hospital	−0.004	−0.003
	(0.017)	(0.015)
Depriviation rank	0.000	−0.002∗
	(0.001)	(0.001)
Scale	−0.002∗	0.001
	(0.001)	(0.001)
Female	0.067	0.212
	(0.116)	(0.120)
Elderly	−0.163	0.166
	(0.269)	(0.275)
CMI	0.045	−0.079
	(0.069)	(0.055)
Constant	0.016	0.105
	(0.090)	(0.068)
Year FE	Yes	Yes
Variance ν_g^P 95% CI	[0.001; 0.004]	[0.001; 0.002]
Intraclass correlation (ICC)	61.79%	46.34%
Model Wald χ^2	26.48	37.14
Observations	641	641
Number of groups	132	132

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
NotRec: Proportion of patients refraining to recommend their PCP. NoAccess: Proportion of patients experiencing access problems at their PCP

Table 5 Estimating model (4) (variance decomposition) using data from the top 60 PCPs that collectively account for 70% of ACS attendances

	(1) $\ln(A_{gt})$	(2) $\ln(A_{gt})$	(3) $\ln(A_{gt})$
Closest hospital		−0.004	−0.004
		(0.021)	(0.024)
Depriviation rank		−0.006∗	−0.006
		(0.003)	(0.003)
Scale			−0.002
			(0.004)
Female			1.245∗
			(0.551)
Elderly			−0.399
			(0.761)
CMI			−0.015
			(0.269)
Constant	−1.604∗∗∗	−1.549∗∗∗	−2.122∗∗∗
	(0.020)	(0.027)	(0.322)
Year FE	Yes	Yes	Yes
$\widehat{\tau}^2$ 95% CI	[0.005; 0.035]	[0.005; 0.032]	[0.005; 0.014]
Intraclass correlation (ICC)	62.90%	61.02%	51.84%
Model Wald χ^2	33.70	34.93	46.66
Observations	293	293	288
Number of groups	60	60	60

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
Note: Number of observation differs due to missing case-mix information in 5 PCP-years.

Table 6 Estimating model (5) (validation) using data from the top 60 PCPs that collectively account for 70% of ACS attendances

	(1) NotRec	(2) NoAccess
$\widehat{u}_g$	0.131	0.215**
	(0.068)	(0.068)
$\widehat{\epsilon}_{gt}$	0.013	0.060
	(0.027)	(0.031)
Closest hospital	−0.009	0.002
	(0.020)	(0.016)
Depriviation rank	0.002	−0.000
	(0.002)	(0.001)
Scale	−0.002	0.002
	(0.002)	(0.001)
Female	0.021	0.214
	(0.135)	(0.118)
Elderly	−0.202	0.252
	(0.475)	(0.358)
CMI	0.125	−0.115
	(0.090)	(0.080)
Constant	−0.047	0.118
	(0.118)	(0.091)
Year FE	Yes	Yes
Variance ν_g^P 95% CI	[0.001; 0.005]	[0.001; 0.002]
Intraclass correlation (ICC)	62.36%	44.87%
Model Wald χ^2	15.88	49.33
Observations	284	284
Number of groups	60	60

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
NotRec: Proportion of patients refraining to recommend their PCP. NoAccess: Proportion of patients experiencing access problems at their PCP

Table 7 **Estimating model (4) (variance decomposition) without closing and expanding PCPs**

	(1) $\ln(A_{gt})$	(2) $\ln(A_{gt})$	(3) $\ln(A_{gt})$
Closest hospital		0.018	0.022
		(0.021)	(0.024)
Depriviation rank		−0.008**	−0.007*
		(0.003)	(0.003)
Scale			−0.004
			(0.004)
Female			0.842
			(0.664)
Elderly			−0.196
			(0.830)
CMI			0.024
			(0.243)
Constant	−1.624***	−1.553***	−1.966***
	(0.018)	(0.028)	(0.359)
Year FE	Yes	Yes	Yes
$\widehat{\tau}^2$ 95% CI	[0.006; 0.026]	[0.005; 0.023]	[0.006; 0.014]
Intraclass correlation (ICC)	46.73%	42.91%	37.89%
Model Wald χ^2	36.39	39.42	43.42
Observations	395	395	390
Number of groups	79	79	79

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
Note: Number of observation differs due to missing case-mix information in 5 PCP-years.

3.2. Excluding PCP closures and expanding PCPs

In the main analysis we include data from PCPs that closed up until one year before the official closure and from PCPs that expanded to new branches up until one year before the expansion. To alleviate concerns that the results may be driven by closures/expansions, we now exclude closing (N=3) and expanding (N=2) PCPs from the entire study period and re-estimate models (1) and (3). Table 7 presents the results form decomposing variation in PCP performance (model (4)) and Table 8 from validating the performance metric using patient surveys (model (5)). The results are quantitatively comparable to the primary analysis presented in the paper.

Table 8 Estimating model (5) (validation) without closing and expanding PCPs

	(1) NotRec	(2) NoAccess
$\widehat{u}_g$	0.119	0.226***
	(0.073)	(0.058)
$\widehat{\epsilon}_{gt}$	0.007	−0.005
	(0.015)	(0.020)
Closest hospital	−0.011	−0.001
	(0.022)	(0.017)
Depriviation rank	0.002	−0.001
	(0.001)	(0.001)
Scale	−0.003*	0.001
	(0.002)	(0.001)
Female	0.046	0.094
	(0.137)	(0.114)
Elderly	−0.158	0.280
	(0.373)	(0.301)
CMI	0.090	−0.116
	(0.083)	(0.062)
Constant	−0.024	0.184*
	(0.112)	(0.074)
Year FE	Yes	Yes
Variance ν_g^P 95% CI	[0.001; 0.005]	[0.001; 0.002]
Intraclass correlation (ICC)	63.46%	47.48%
Model Wald χ^2	18.22	35.85
Observations	386	386
Number of groups	79	79

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
NotRec: Proportion of patients refraining to recommend their PCP. NoAccess: Proportion of patients experiencing access problems at their PCP

Table 9 Validation of the performance measure via the survey's second annual measurement

	(1) NotRec	(2) NoAccess
$\widehat{u}_g$	0.230∗	0.203∗∗
	(0.090)	(0.071)
$\widehat{\epsilon}_{gt}$	0.009	−0.008
	(0.023)	(0.027)
Closest hospital	0.014	0.009
	(0.018)	(0.016)
Depriviation rank	0.000	−0.001
	(0.002)	(0.001)
Scale	−0.002	0.002
	(0.002)	(0.001)
Female	0.514	−0.054
	(0.279)	(0.242)
Elderly	0.420	0.209
	(0.412)	(0.324)
CMI	−0.134	−0.151
	(0.143)	(0.105)
Constant	−0.040	0.296∗
	(0.215)	(0.143)
Year FE	Yes	Yes
Variance ν_g^P 95% CI	[0.001; 0.005]	[0.001; 0.002]
Intraclass correlation (ICC)	61.16%	41.36%
Model Wald χ^2	17.22	18.26
Observations	244	244
Number of groups	82	82

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
NotRec: Proportion of patients refraining to recommend their PCP. NoAccess: Proportion of patients experiencing access problems at their PCP

3.3. Second annual measurement moment for patient survey

The patient survey data that we use for validating the performance metric is reported biannually (July/June and December) up until 2016 and annually from thereon (July). In the primary analysis, we relied on the first measurement (July/June). In this section we replicate this validation step with the second measurement (December). Note that this means that we have fewer annual observations (2013-2015). Table 9 presents the results and again shows that the PCP's deviation from the benchmark $\widehat{u}_g$ is positively associated with patients' responses to the survey.

4. Bootstrapping the effect size

In §3.5 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 PCPs with replacement.
2. We estimate model (1) based on this new dataset.

3. We use the estimated model to calculate the log-transformed proportion of ACS-attendances for each PCP-year: $\widehat{ln(A)}_{gt} = \widehat{\beta}_0 + \widehat{\beta}_C C_{g,t} + \widehat{u}_g$.

4. To derive the untransformed ACS-proportion we calculate $\widehat{A}_{gt} = \exp(\widehat{ln(A)}_{gt}) \times \exp(\widehat{\sigma}^2/2)$.

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

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

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

8. 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 already perform at the 25% percentile are left unchanged.

9. We repeat steps 3 to 6 to obtain the predicted number of ACS attendances for the counterfactual scenario.

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

11. We repeat steps 1 to 10 (10,000 times) to derive a 95% confidence interval for the difference in ACS attendances based on the empirical distribution.

5. Alternative performance indicator: ACS admission rates

In §4.3 of the main paper we argued that patient ACS attendances at the ED is a better measure of PCP performance compared to ACS admissions. In this section, we assess how far a performance measure based on admissions is able to differentiate between PCPs.

For every PCP g we denote the count ACS admissions in period t by:

$$\tilde{P}_{gt} = \sum_{i=1}^{N} X_{i,g,ACS,t},$$

where $X_{i,g,c,t}$ is equal to 1 if patient $i \in \{1, ..., N\}$ is registered with practice $g \in \{1, ..., G\}$ in period t and is admitted to the hospital with a condition c that is ACS and 0 otherwise.

Similar to the primary analysis, we normalize this measure by dividing with all hospital admissions originating from PCP g in period t:

$$\tilde{A}_{gt} = \frac{P_{gt}}{\sum_{c' \in C} \sum_{i=1}^{N} X_{i,g,c',t}},$$

where $X_{i,g,c,t}$ is as defined above and c' includes patients with ACS and non-ACS conditions. A histogram of $\tilde{A}_{gt}$ is shown in Figure 3.

The proportion of ACS admissions $\tilde{A}_{gt}$ is then decomposed in a manner similar to the primary analysis:

$$ln(\tilde{A}_{gt}) = \alpha_0 + \alpha_C C_{gt} + \epsilon_{gt}. \tag{6}$$

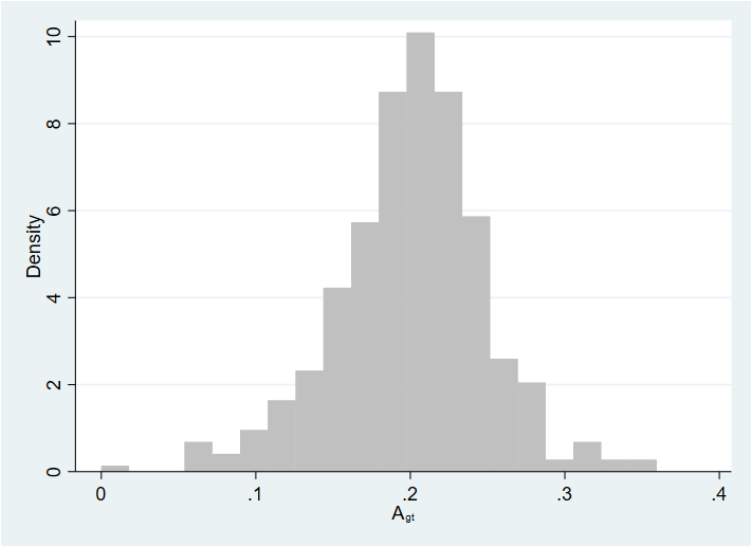


Figure 3 Histogram ACS admission rates

The results of model (6) are presented in Table 10. Column (1) presents the results of a model without any controls except for the year fixed effects. Column (2) controls for PCP location and deprivation. Column (3) adds controls for PCP scale, proportion of female and elderly patients, and CMI. Adding the controls only marginally explains some of the between PCP variance as the ICC decreases from 27.18% in the model without controls (Column 1) to 23.40% in the most detailed model (Column 3). The most detailed model suggest that less than one quarter of the variance in PCP performance is systematic at the PCP level. To put this in perspective, the model based on ACS ED attendances suggest that more than one third is systematic at the PCP level. The correlation between an indicator based on admissions and an indicator based on attendances is positive (0.567, $p<0.001$), which suggests a partial measurement overlap. However, measuring performance on the basis of ACS attendances attributes more performance differences to the PCP level, whereas an measurements based on ACS admissions attributes more performance differences to stochastic fluctuations.

We then assess whether the performance measure based on ACS admissions is correlated with patient surveys. We use the same approach as in the primary analysis and present the results in table 11. We find that the PCP's performance (the random effect $\widehat{u}_g$), as measured by ACS admissions is not associated with the survey measures P_{gt}.

In conclusion, if one has to rely on ACS admission data, it is more challenging to differentiate between PCP practices because such a measure exhibits less between variance. In addition, this measure does not show evidence of external validity.

Table 10 Decomposing variation in PCP performance based on ACS admissions

	(1) $ln(\tilde{A}_{gt})$	(2) $ln(\tilde{A}_{gt})$	(3) $ln(\tilde{A}_{gt})$
Closest hospital		0.009	0.010
		(0.007)	(0.007)
Depriviation rank		−0.001	−0.000
		(0.001)	(0.001)
Scale			−0.002
			(0.001)
Female			0.108
			(0.067)
Elderly			0.219
			(0.217)
CMI			0.005
			(0.062)
Constant	0.208***	0.212***	0.154*
	(0.006)	(0.007)	(0.060)
Year FE	Yes	Yes	Yes
$\widehat{\tau}^2$	0.001	0.001	0.001
$\widehat{\tau}^2$ 95% CI	[0.000; 0.001]	[0.000; 0.001]	[0.006; 0.014]
Intraclass correlation (ICC)	27.18%	26.65%	23.40%
Model Wald χ^2	19.70	22.56	38.15
Observations	408	408	401
Number of groups	84	84	83

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
Note: Number of observation differs due to missing case-mix information in 7 PCP-years.

Table 11 Validation of a performance measure based on ACS admissions

	(1) NotRec	(2) NoAccess
$\widehat{u}_g$	−0.010	0.285
	(0.386)	(0.338)
$\widehat{\epsilon}_{gt}$	0.003	−0.056
	(0.045)	(0.055)
Closest hospital	−0.005	0.001
	(0.020)	(0.017)
Depriviation rank	0.002	−0.001
	(0.002)	(0.001)
Scale	−0.003	0.001
	(0.002)	(0.001)
Female	0.037	0.038
	(0.133)	(0.196)
Elderly	−0.142	0.219
	(0.366)	(0.333)
CMI	0.084	−0.092
	(0.083)	(0.063)
Constant	−0.015	0.189
	(0.109)	(0.101)
Year FE	Yes	Yes
Variance ν_g^P	0.003	0.002
Variance ν_g^P 95% CI	[0.001; 0.005]	[0.001; 0.003]
Intraclass correlation (ICC)	63.49%	51.02%
Model Wald χ^2	15.89	14.53
Observations	397	397
Number of groups	83	83

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
NotRec: Proportion of patients refraining to recommend their PCP. NoAccess: Proportion of patients experiencing access problems at their PCP

6. ED-sensitive conditions

Table 12 provides a list of the ED-sensitive conditions used in the placebo test of §4.2 in the main paper.

Table 12 ED-sensitive conditions used in the placebo test

Abdominal/ Thoracic Aneurysm
Acute Coronary Syndrome
Acute Pancreatitis
Acute Renal Failure
Alcohol Intoxication/ Overdose / Withdrawal
Anaphylactic Shock
Appendicitis
Arrhythmia
Arterial Embolism / Thrombosis
Burn (Full-thickness)
Carbon Monoxide Poisoning
Cardiac Arrest
Cardiogenic Shock
Cardiomyopathy
Chemical Burn/ Exposure
Chest Pain
Cholecystitis
Complication Following Abortion
Ectopic Pregnancy
Electric Shock
Electrical Burn
GI Haemorrhage
Gunshot/stab wound
Haemopneumothorax
Heat Stroke / Sun Stroke
Hepatic Failure Or Coma
Hypothermia
Infective/Non-infective Complication Contraceptive Device
Intracerebral Haemorrhage
Mechanical/Non-mechanical Complication of Cardiac Electronic Device (inc PPM, ICD)
Medication/Drug Overdose/ Poisoning
Meningitis
(Meningo) Encephalitis
Myocardial Infarct
Myocardial Ischaemia
Myocarditis
Oesophageal Perforation
Open Chest Wound
Open Fracture
Other Cardiac Condition
Pancreatitis
Pericardial Effusion
Pericarditis
Peritonitis
Pneumothorax
Post Op Wound Complication
Pulmonary Embolism
Radiation Burn/ Injury
Respiratory Arrest/ Failure
Sepsis
Stroke
Subarachnoid Haemorrhage
Supraventricular Tachycardia
Surgical Procedure Complication
Transient Ischaemic Attack
Ventriculoperitoneal Shunt Complication

7. The performance decomposition methodology vs patient surveys

Given that patient surveys are administered regularly and the results are publicly available, one might wonder whether the survey outcomes could be used directly to identify best operational practices. In the main paper we argue that there are multiple reasons in favour of the performance decomposition methodology. In this section we replicate the analysis of §5 of the main paper, where instead of using the PCP random effect $\hat{u}_g$ and the dependent variable we instead use the survey outcome:

$$P^i_{gt} = \delta_0 + \delta_1 PpP_{gt} + \delta_2 PpP_{gt} \times Scale_{gt} + \delta_3 Scale + \delta_C C_{gt} + \nu_g + \epsilon^P_{gt}, \tag{7}$$

The results are presented in Table 13 and they fail to detect that the number of patients per FTE is a driver that affects PCP quality (δ_1 is not statistically different to zero in all specifications). The results also do not hint at economies of scale, there is no statistical indication that practice scale moderates the relationship between patient-to-staff ratio and performance (δ_2 is not statistically different to zero in all specifications). Compared to the performance decomposition methodology, which exhibites less uncertainty around the effect estimates, using surveys indeed makes it more challenging to identify drivers of performance in primary care.

Table 13 Relationship between patient survey measures and patient-to-staff ratio

	(1) NotRec	(2) NotRec	(3) NotRec	(4) NoAccess	(5) NoAccess	(6) NoAccess
PpP	0.004	0.004	0.001	0.004	0.004	0.009
	(0.003)	(0.003)	(0.007)	(0.003)	(0.003)	(0.006)
Scale		−0.003	−0.004		0.001	0.003
		(0.002)	(0.002)		(0.001)	(0.002)
PpP × Scale			0.000			−0.001
			(0.001)			(0.001)
Closest hospital		−0.002	−0.003		−0.000	0.000
		(0.020)	(0.019)		(0.017)	(0.017)
Depriviation rank		0.002	0.002		−0.001	−0.001
		(0.002)	(0.002)		(0.001)	(0.001)
Female		0.092	0.093		0.056	0.054
		(0.124)	(0.123)		(0.187)	(0.182)
Elderly		−0.060	−0.083		0.262	0.290
		(0.427)	(0.430)		(0.372)	(0.372)
CMI		0.024	0.030		−0.092	−0.099
		(0.130)	(0.131)		(0.107)	(0.109)
Constant	0.070***	0.010	0.010	0.122***	0.173	0.170
	(0.007)	(0.147)	(0.147)	(0.008)	(0.122)	(0.120)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Variance ν^P_g 95% CI	[0.001; 0.005]	[0.001; 0.005]	[0.001; 0.005]	[0.001; 0.003]	[0.001; 0.003]	[0.001; 0.003]
Intraclass Correlation (ICC)	64.54%	63.30%	63.43%	49.48%	49.61%	49.52%
Model Wald χ^2	11.81	17.37	20.53	9.06	12.40	12.18
Observations	369	369	369	369	369	369
Number of groups	83	83	83	83	83	83

Clustered standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05.
NotRec: Proportion of patients refraining to recommend their PCP.
NoAccess: Proportion of patients experiencing access problems at their PCP.

References

Mundlak Y (1978) On the pooling of time series and cross section data. *Econometrica* 46(1):69–85.

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