Supplementary Material
Economies of Scale and Scope in Hospitals

Michael Freeman
Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom mef35@cam.ac.uk

Nicos Savva
London Business School, Regent's Park, London NW1 4SA, United Kingdom nsavva@london.edu

Stefan Scholtes
Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom s.scholtes@jbs.cam.ac.uk

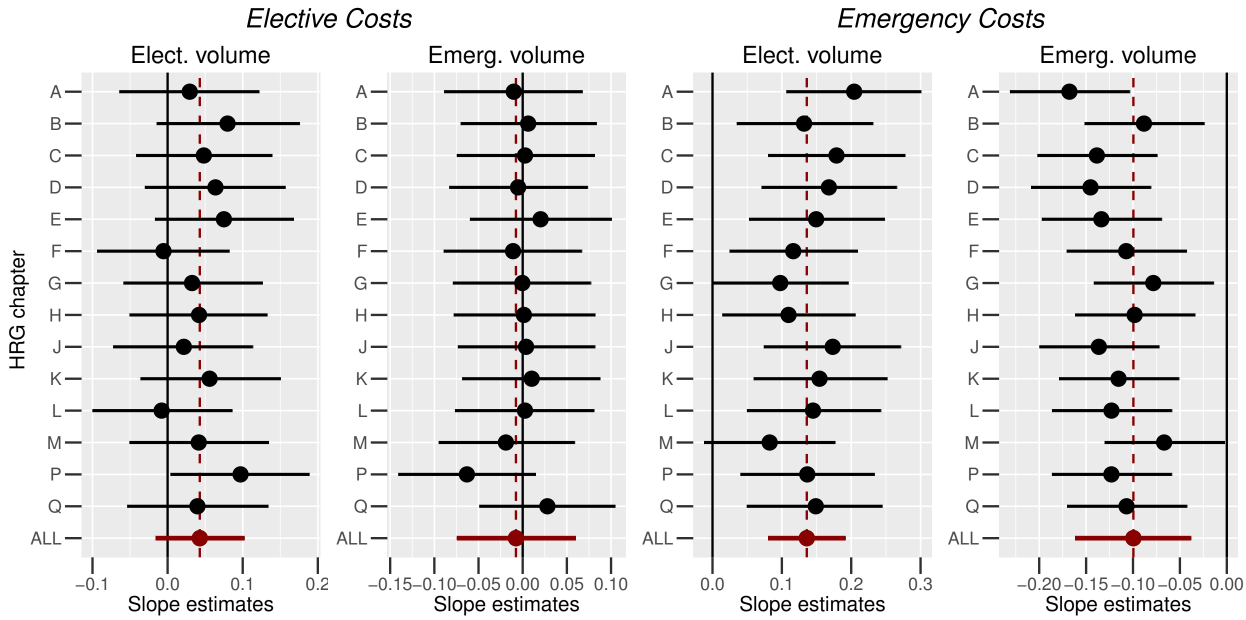
1. Introduction

This appendix contains supporting material designed to accompany the investigation presented in the main paper. In §2 we provide random slope estimates for the effects of volume from *other* service lines, to augment those provided for volume of the *same* service line given in the §6.1 of the paper. In §3 we show that there is no evidence that the errors are autocorrelated. In §4 we investigate the possibility of non-linear volume effects, and find little evidence to suggest this is the case. In §5 we discuss the fact that elective service lines might be formed endogeneously based on financial viability, and show how we account for this, provide additional robustness, and discuss how – if anything – this would be expected to work against our findings. In §6 we report on the results a number of additional tests that (i) are performed on a subset of data corresponding to hospital trusts that are more geographically isolated, (ii) limit the possibility of extreme cost outliers driving the results, (iii) compare hospital trusts based on a set of common (rather than all) HRGs that are performed by most (>80%) of trusts, (iv) re-run the models on a subset of the service-lines for which hospital trusts treat a high enough volume of patients, and (v) restrict the sample to trusts that operate only a single hospital site. The results from all of the models in §6 are in line with those reported in the paper.

2. Random slopes – hospital trust volume effects

In §6.1 of the paper we report on random slopes estimates for the effect of same–service-line volume on hospital trust costs. In addition to this, it is interesting to consider whether the effect of volume from the other service lines on cost of the focal service line differs by service line. To do this we have estimated random slope estimates for the corresponding variables replacing the coefficients β_2^h and β_4^h in equation (7) in the paper with random, service-line dependent slopes: $\beta_{2,(C)[i]}^h$ and $\beta_{4,(C)[i]}^h$. As

Figure 1 Random slope coefficient estimates for the effect of volume from the other service lines on the cost of the focal service line, reported by service line (black) and combined (red), with bootstrapped 95% confidence intervals.



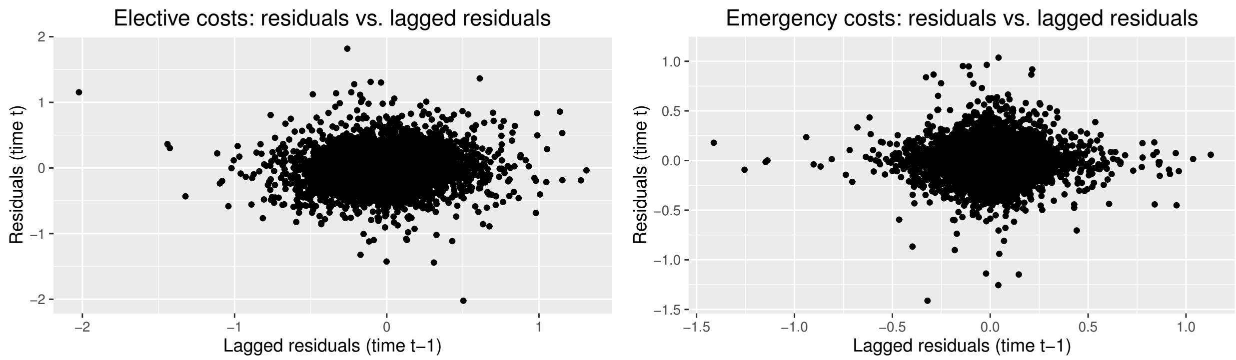
Note. A - Nervous system; B - Eyes and periorbita; C - Mouth, head, neck and ears; D - Respiratory system; E - Cardiac surgery and primary cardiac conditions; F - Digestive system; G - Hepatobiliary and pancreatic system; H - Musculoskeletal system; J - Skin, breast and burns; K - Endocrine and metabolic system; L - Urinary tract and male reproductive system; M - Female reproductive system; P - Diseases of childhood and neonates; Q - Vascular system.

in the paper, to allow the random slopes to be correlated with the service-line specific intercepts we need to also replace the service-line fixed effect in equation (8) in the paper with the random effect, $\alpha_{(C)[i]}$. We then model $(\alpha_{(C)[i]}, \beta^h_{2,(C)[i]}, \beta^h_{4,(C)[i]})$ using a trivariate normal distribution, to allow for correlation between the random effects. The results are plotted – together with bootstrapped 95% confidence intervals using 10,000 simulations from the posterior distribution of the MLMs – in Figure 1. We have also plotted the combined slope estimates from the main estimations and comparing against this it can be seen that the direction of the individual effects are consistent with the combined estimates, with 95% confidence intervals overlapping in nearly all cases.

3. Autocorrelated errors

One concern when working with a panel of time series data is that errors may be autocorrelated, i.e. costs change slowly and e.g. high costs in one year may indicate that costs will be high in the next year, also. The standard errors are often underestimated when autocorrelation of the error terms (at low lags) are positive. This is unlikely to be a major issue for our analysis, since results are highly significant and standard errors would have to be vastly underestimated for the results to be misidentified. A bigger concern, however, is that autocorrelation of the errors may bias the coefficient estimates in the within-between model formulation. We investigate this further in this

Figure 2 **Plots of residuals (time t) against lagged residuals (time $t-1$) for elective costs (left) and emergency costs (right).**



section.

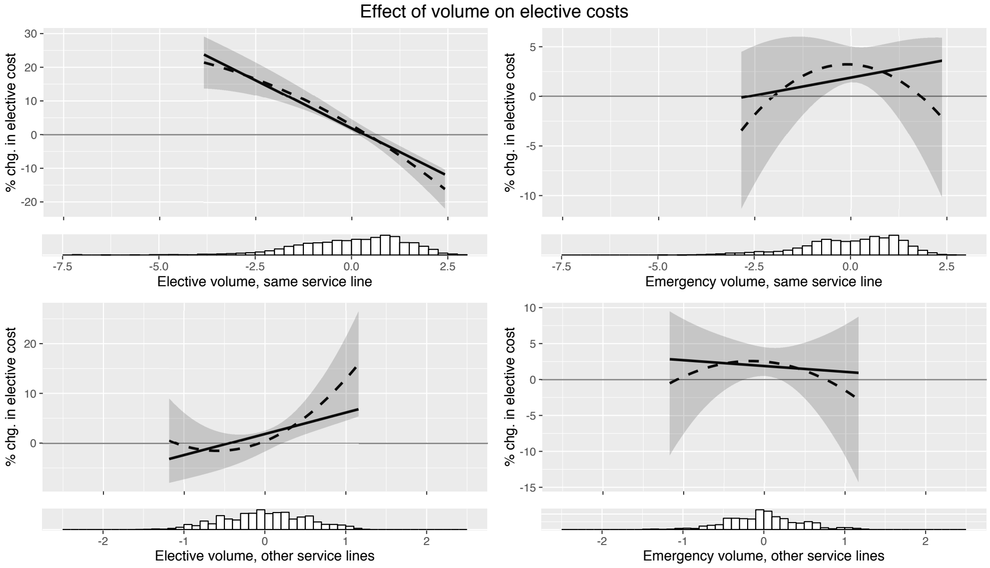
To do this, we have taken two approaches. In the first, we regress (using OLS) the residuals (at time t) from the within-between multilevel models (MLMs) against the lagged residuals (at time $t-1$). A plot of residuals vs. lagged residuals is provided in Figure 2, showing little evidence of any correlation and hence suggesting that our models account for much of the within-trust and time-related correlation of the errors. This is confirmed by OLS models, with only $\sim 1.4\%$ of the variance in the residuals for elective costs explained by the lagged residuals, and $< 0.1\%$ for emergency costs.

We follow the informal approach described above with a traditional testing method. The standard test for the presence of first-order correlation is the Durbin-Watson statistic. However, this test can only be performed if the panel is balanced. For an unbalanced panel the recommended approach is to instead calculate the Baltagi-Wu locally best invariant (LBI) test statistic (Baltagi and Wu 1999). We estimate this using the `xtregar` command in Stata 12.1. Note that the models that we estimate this statistic for are not identical to those presented in the paper (since the particular command in Stata does not allow the estimation of multiple random effects, and so instead we are only able to include trust–service-line REs), but is sufficiently similar (and, if anything, since the control structure in the paper includes additional time-related controls, the estimates reported here are likely to be conservative). Doing so we calculate the LBI statistic to be 1.82 for elective costs and 1.85 for emergency costs. While critical values are not available in Baltagi and Wu (1999), if there were no evidence of first-order autocorrelation then these should take value 2. Since the LBI statistics are close to 2 in value, we conclude there is little reason to be concerned about autocorrelated errors.

4. Non-linear volume effects

In the paper we assume the effects of (log) volume on (log) cost is linear, i.e. a 1% increase in volume has an $x\%$ effect on cost, regardless of the initial level of volume. Here we discuss relaxing

Figure 3 Plots of estimated (mean-centered) volume effects on elective costs in models with only linear volume effects (solid lines) and in models also with non-linear volume effects (dashed lines).



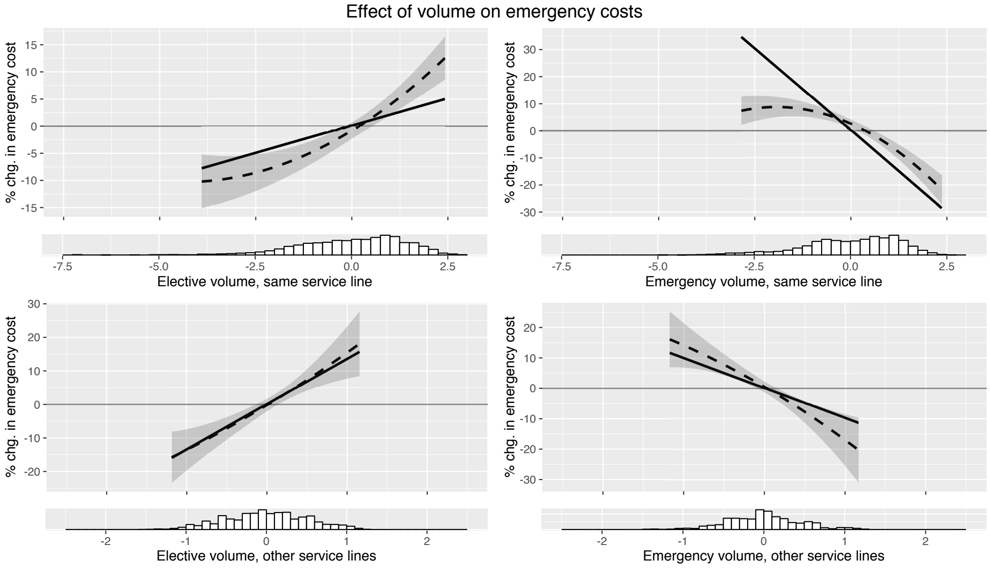
this assumption to allow for non-linear volume effects. We do this by including the squared values of the between–hospital-trust volume terms in the main multilevel models.

In Figures 3 and 4 are plotted for the elective and emergency patient types, respectively, the estimated between-effects of volume in models with linear only volume effects (i.e. the estimated effects reported in the paper) and in models with the inclusion of non-linear (squared terms) volume between-effects. 95% confidence bands for the non-linear effects are also plotted. As shown, with the exception of the effect of same–service-line emergency volume on emergency costs (see top-right of Figure 4), there is little evidence that the interpretation of the results would change significantly if we had instead used a non-linear formation of volume.

For the effect of same–service-line emergency volume on emergency costs, the discrepancy between the non-linear and the linear estimations deserves further investigation. One possibility is that the non-linear effects identified may in part by driven by extreme points / outliers. To investigate this further, we have re-run the estimations using a subset of the data for which the relevant volume (shown in the histogram in the top-right of Figure 4) is constrained to take values between -2.5 and 2.5. After doing so and re-estimating, we find no further evidence of non-linearity, suggesting that the non-linear effects identified in Figure 4 are most likely driven by the presence of these outliers. Returning to the linear formulation of this model and re-estimating the effect of same–service-line emergency volume on emergency cost, we find this to equal -0.051 (p-value < 0.001), as compared to -0.121. We note this difference in scale in §6.3 of the paper.

While the other non-linear estimations do not indicate strong sensitivity of the results to outliers,

Figure 4 **Plots of estimated (mean-centered) volume effects on emergency costs in models with only linear volume effects (solid lines) and in models also with non-linear volume effects (dashed lines).**



in §6.4 of this supplementary material document we also report results where volume is constrained to be above a specified minimum level.

5. Endogenous service line composition

Not every hospital trust may offer every type of treatment, and while hospitals in the UK are not as financially driven as their US counterparts, the choice of which treatments to offer (i.e. the composition of the service lines) might still be related to the financial viability of different treatment options. To determine the extent to which hospital trusts select to treat particular types of patients or conditions, we can re-use the weights which were employed to case-mix adjust costs and LOS in the paper. Specifically, we defined weights $\alpha_{tcp} = \frac{n_{tcp}}{\sum_{c \in C} n_{tcp}}$, where n_{tcp} is the total number of patients of type p with HRG c in year t across all trusts. In each hospital trust we can then define a new variable

$$\mathbf{Prop}_{thCp} = \sum_{c \in C_{thp}} \alpha_{tcp}\,, \tag{1}$$

where C_{thp} is the subset of HRGs c in service line C for patients of type p that are observed in trust h in year t. Then (1) gives the proportion of the 'average' elective ($propEl$) and emergency ($propEm$) patient case-mix that a hospital trust treats in that service line in that year. A plot of these proportions (for each of the service lines) is given in Figures 5 and 6 for the elective and emergency patient types, respectively.

As can be seen in Figure 5 there is some evidence that not all elective treatments are offered at all hospital trusts, while Figure 6 shows that – other than for Chapter B, which relates to conditions

Figure 5 Proportion of the "average" elective case-mix offered in each service line by every trust in any year.

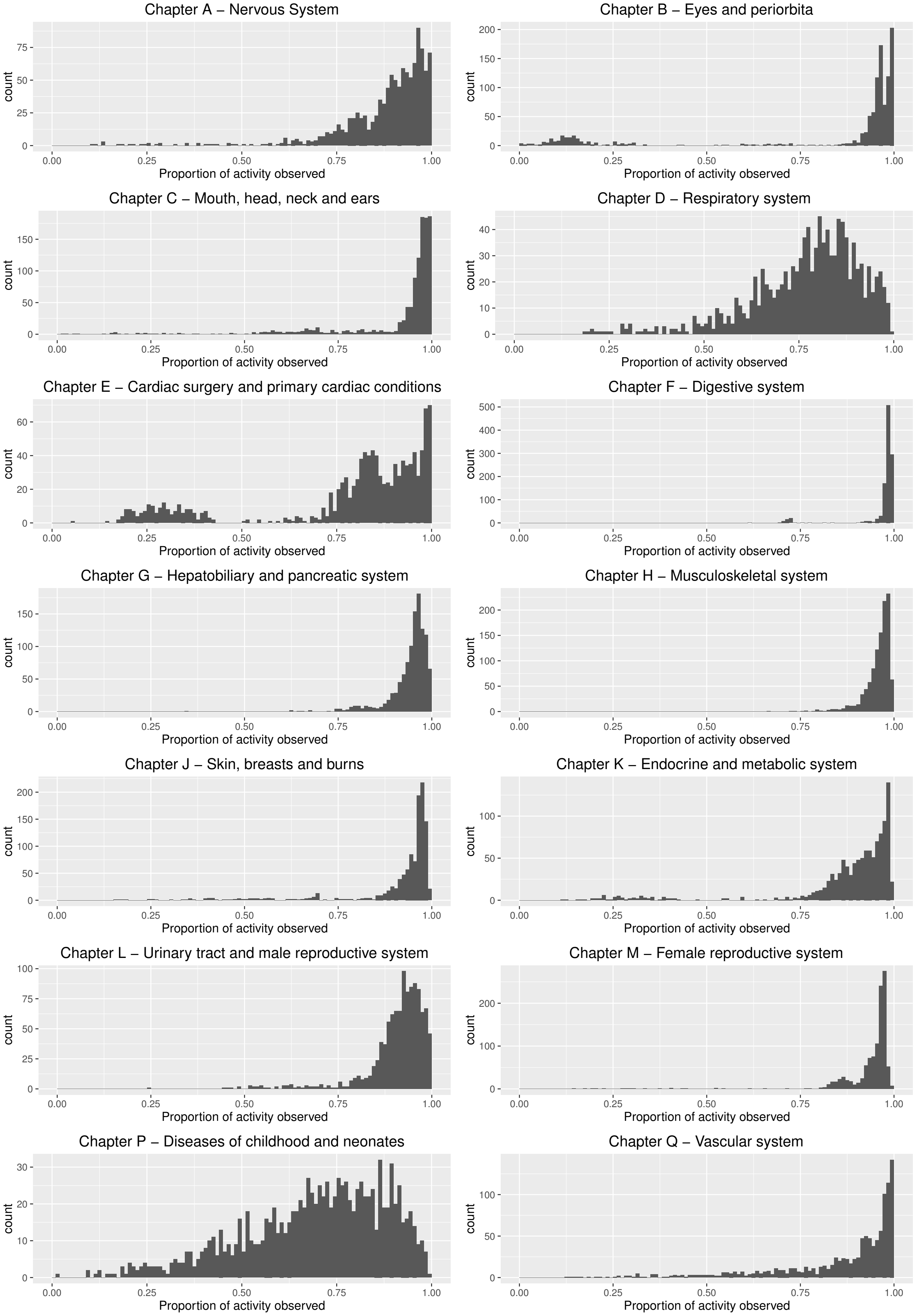


Figure 6 Proportion of the "average" emergency case-mix offered in each service line by every trust in any year.

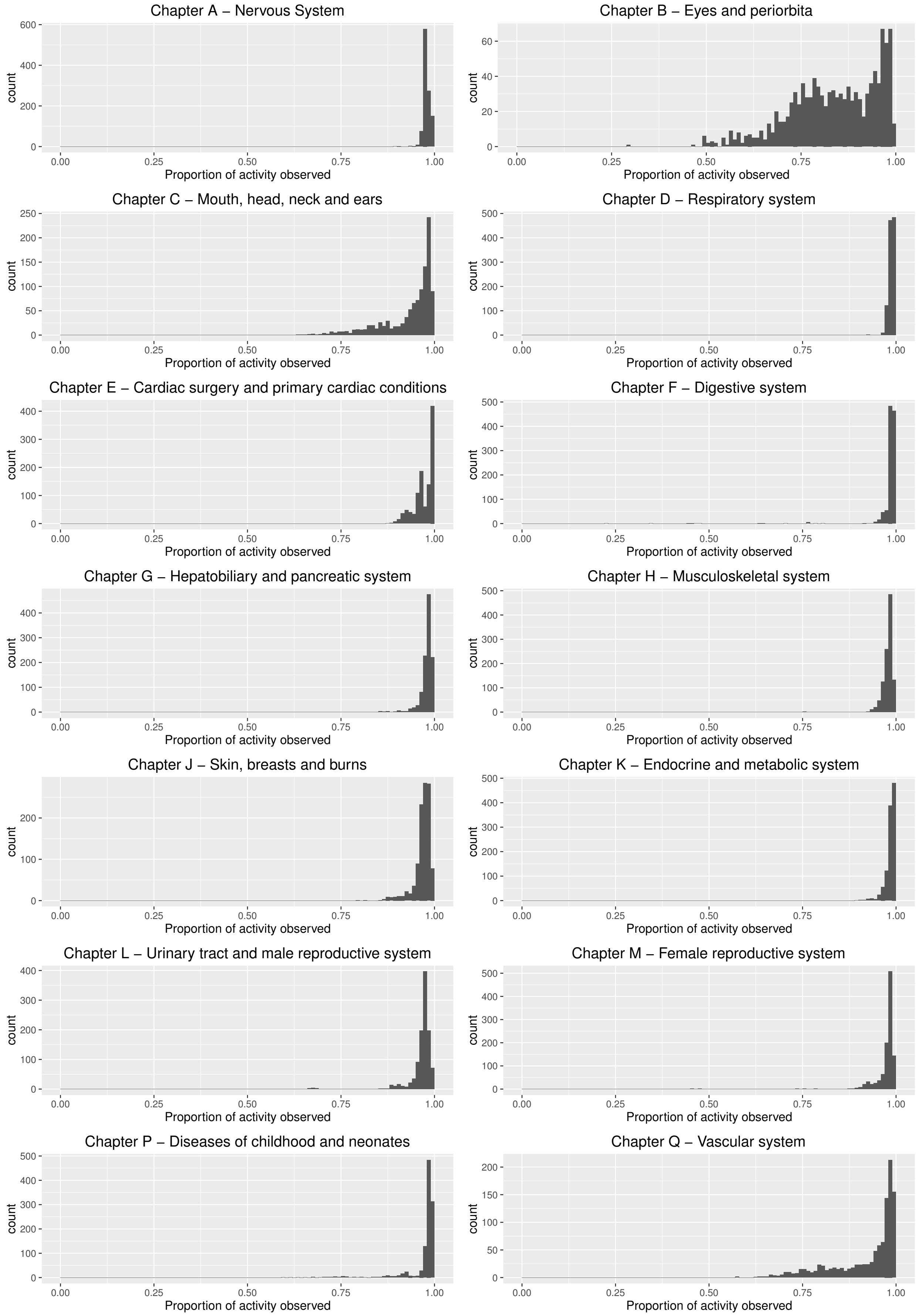


Table 1 Estimated effect of proportion of HRGs treated within a service-line

	Costs		Length of stay	
	Elective	Emergency	Elective	Emergency
Within effects				
Prop. elect.	0.177^{***}	–	0.144^{***}	–
	(0.028)	–	(0.014)	–
Prop. emerg.	–	0.104^{*}	–	-0.209^{***}
	–	(0.045)	–	(0.029)
Between effects				
Prop. elect.	0.465^{***}	–	0.017	–
	(0.055)	–	(0.024)	–
Prop. emerg.	–	1.221^{***}	–	0.427^{***}
	–	(0.083)	–	(0.059)

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$; Coefficient estimates and standard errors reported for the main effects.

of the eyes and periorbita – there is little evidence of emergency treatments not being offered at all trusts (unsurprising, as the unpredictable nature of patient arrivals to the ED means that hospitals have little choice over which emergency patients they will serve). For elective service lines, though, it is possible that the service lines are formed endogenously, i.e. hospital trusts may choose to only offer treatment to more profitable types of patients. To account for this in the paper we:

1. Construct the dependent variable by dividing actual costs by the 'average' cost, with both calculated using the same weights (i.e. the same case-mix). So, if e.g. only 80% of the HRGs in a service-line appear in the numerator, then only the same 80% of HRGs will appear in the denominator also. In this way costs are adjusted for observable differences in the service offering.

2. Use hospital trust and/or trust–service-line fixed- and/or random-effects, to capture systematic, time-invariant differences in the costs at different trusts due to e.g. unobservable differences in the types of treatment offered.

3. Control in the elective costs models with *propEl*, and in the emergency costs models with *propEm*, to capture the fact that costs may differ depending on the diversity of services offered within a service line. (In the within-between models we instead control with the hospital trust mean-centered proportions and mean proportions, i.e. $propEl - \overline{propEl}$ and $\overline{propEl}$, and similarly for *propEm*.)

The estimated coefficients for *propEl* and *propEm* in the within-between models are given in Table 1. As is shown, the greater the proportion of services offered within a particular service line (i.e. the wider the scope and more varied the service offering), the greater also the cost. This indicates that, if anything, those hospital trusts that choose to offer a narrow range of services, and hence are likely to have lower volume (as indicated by the strong positive correlation of 0.77 between $\overline{propEl}$ and $\overline{ln(nElServ)}$, and of 0.67 between $\overline{propEm}$ and $\overline{ln(nEmServ)}$), tend also to offer those less costly services. This would serve to work in the *opposite* direction to the same–service-line, same–patient-type volume effects that we find in the paper, and if this were driving

Table 2 **Model parameter estimates - with service-line:proportion interactions as additional controls**

	Costs		Length of stay	
	Elective	Emergency	Elective	Emergency
Between effects				
Elect. vol. (focal SL)	-0.055^{***}	0.017^{***}	-0.026^{***}	0.009^{**}
	(0.010)	(0.004)	(0.004)	(0.003)
Emerg. vol. (focal SL)	0.009	-0.137^{***}	0.017^{**}	-0.089^{***}
	(0.014)	(0.011)	(0.006)	(0.008)
Elect. vol. (other SLs)	0.058	0.134^{***}	-0.001	0.064^{*}
	(0.031)	(0.028)	(0.014)	(0.026)
Emerg. vol. (other SLs)	-0.013	-0.084^{**}	0.018	-0.036
	(0.034)	(0.031)	(0.016)	(0.029)
Model fit				
Observations	15,339	15,354	15,339	15,354
Marginal R^2	0.114	0.190	0.125	0.110
Conditional R^2	0.519	0.624	0.468	0.744
Bayesian inf. crit.	559.6	-11,074.3	-20,496.8	-23,867.4

${}^{*}p < 0.05$; ${}^{**}p < 0.01$; ${}^{***}p < 0.001$; Coefficient estimates and standard errors reported for the between effects; Marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

our results we would expect instead that the greater the volume of activity of the same-type in the same service line, the wider the service offering, and hence the higher the cost.

To confirm the robustness of the results, we have re-run the analysis in the paper under two alternative scenarios. In the first, we recognize that – as per Figure 5 – in the various elective service lines the degrees to which hospital trusts appear to restrict the scope of their services differ. For example, the vast majority of hospital trusts offer the full range of musculoskeletal services (Chapter H), while this is far more varied for diseases of the respiratory system (Chapter D). Therefore, we repeat the analysis in the paper by including $propEl$ (and $propEm$ also, for completeness) interacted by the service line indicator, to allow the coefficients to vary depending on the service line. The results (between effects reported only, for brevity) are presented in Table 2. Comparing the MLM model results to those in the paper we find them to be consistent.

In addition to the robustness check above, as a second check we also re-run the analysis for a subset of the data in which we only include observations for which (i) $propEl, propEm > 0.5$, i.e. the hospital trust treats at least 50% of the expected case-mix, and (ii) $\overline{propEl}, \overline{propEm} < 0.8$, i.e. on average over the years observed the hospital trust treats at least 80% of the expected case-mix. This reduces the sample by 17.0% for the elective patient type, and 4.2% for the emergency patient type. In this model we do not control for $propEl$, $propEm$, $\overline{propEl}$ or $\overline{propEm}$. This alleviates concerns that the results in the paper may be spurious and caused by the high correlation between $\overline{propEl}$ and $\overline{ln(nElServ)}$, and between $\overline{propEm}$ and $\overline{ln(nEmServ)}$. The findings (between effects reported only, for brevity) are reported in Table 3. Again, results are consistent with those documented in the paper.

Table 3 Model parameter estimates - excluding observations where low proportion of case-mix treated

	Costs		Length of stay	
	Elective	Emergency	Elective	Emergency
Between effects				
Elect. vol. (focal SL)	-0.046^{***}	0.040^{***}	-0.030^{***}	0.016^{***}
	(0.010)	(0.005)	(0.004)	(0.003)
Emerg. vol. (focal SL)	0.033^{*}	-0.077^{***}	0.019^{**}	-0.069^{***}
	(0.016)	(0.010)	(0.007)	(0.007)
Elect. vol. (other SLs)	0.061^{*}	0.124^{***}	0.014	0.063^{*}
	(0.030)	(0.028)	(0.014)	(0.026)
Emerg. vol. (other SLs)	-0.035	-0.123^{***}	0.011	-0.061^{*}
	(0.035)	(0.031)	(0.016)	(0.029)
Model fit				
Observations	12,732	14,708	12,732	14,708
Marginal R^2	0.095	0.151	0.117	0.103
Conditional R^2	0.561	0.627	0.561	0.766
Bayesian inf. crit.	-3,020.6	-11,496.5	-22,406.0	-24,448.2

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; Coefficient estimates and standard errors reported for the between effects; Marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

Overall, we have demonstrated that while endogenous service line formation is a valid concern, we have taken steps to alleviate this in our models through an appropriate control structure, and have discussed that, if anything, this should work against the effects we find since larger hospital trusts tend to offer more varied services and treat a wider range of conditions, which we show has the effect of increasing their costs. We have also fit two further models that (i) added additional controls to account for the fact that the potential endogeneous effect might be service line specific, and (ii) restricted the sample to only those hospital trusts that offer a wide range of services. Results in both of these cases are consistent with those in the paper.

6. Modeling - data alternatives

In this section we report the results from a number of other estimations made using

6.1. Geographically dispersed hospital trusts

In Table 4 we report the within-effects estimated for a subset of hospital trusts constrained to be 20km apart (see §6.2 of the paper for details). The main results are comparable in sign and scale to those reported in the paper, though the significant reduction in sample size (a 63% decrease in trust-year observations from 1,097 to 406, and of observations in general from ∼15,400 to ∼5,700) means that significance has reduced (and may in some cases results no longer appear statistically significant, e.g. the effect of volume from other service lines on emergency costs).

6.2. Capped costs

It is not uncommon for hospital trusts' costs to be magnified (or shrunk) by a few extremely expensive (or low-cost) patients. Therefore, when government agencies calculate hospital trust compensation based on HRG tariffs, the costs are often trimmed to exclude extreme observations.

Table 4 Model parameter estimates - subset of geographically dispersed hospitals

	Costs		Length of stay	
	Elective	Emergency	Elective	Emergency
Between effects				
Elect. vol. (focal SL)	-0.054^{**}	0.023^{**}	-0.042^{***}	0.006
	(0.019)	(0.009)	(0.008)	(0.007)
Emerg. vol. (focal SL)	0.040	-0.119^{***}	0.036^{**}	-0.079^{***}
	(0.027)	(0.021)	(0.012)	(0.015)
Elect. vol. (other SLs)	0.011	0.116	0.025	0.101
	(0.058)	(0.061)	(0.026)	(0.058)
Emerg. vol. (other SLs)	−0.026	−0.051	−0.007	−0.089
	(0.065)	(0.066)	(0.029)	(0.062)
Model fit				
Observations	5,683	5,684	5,683	5,684
Marginal R^2	0.077	0.107	0.101	0.140
Conditional R^2	0.474	0.567	0.449	0.767
Bayesian inf. crit.	415.5	-3,184.1	-8,108.3	-8,665.6

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; Coefficient estimates and standard errors reported for the between effects; Marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

We adopt a similar approach by limiting the influence of "extreme" costs by capping them at a minimum or a maximum value. We do this by constraining in every hospital trust h the average cost of treating patients with HRG c and of patient type p (i.e. cost_{thcp}) to take maximum value equal to K multiplied by the across-trust median in that year t, and minimum value equal to $1/K$ multiplied by the across-trust median. This leave the same sample and number of observations as originally available, but limits the extent to which extreme values for individual cost can affect the results.

We report in Table 5 the 'between-effects' of the cost estimations where K is set equal to 10 (Columns 2 and 3) and 5 (Columns 4 and 5). The results are nearly identical to those reported in the paper, suggesting little evidence that they are overly influenced by the presence of extreme (i.e. very high or very low cost) observations.

6.3. Common HRGs

In the paper we compared hospital trusts across the set of all HRGs c treated in year t. An alternative to this would have been to compare hospital trusts across a set of common HRGs, i.e. excluding those conditions that are more rare for which treatment is typically only offered in large, teaching or specialist hospitals. This has the additional benefit that it partially reduces the potential for bias caused by endogeneous service line formation (see §5 of this document), since we only compare costs against a base set of HRGs c that are widely provided (and so typically higher volume also, with less discretion in their provision).

To achieve this, we specify that an HRG is only included in the comparison if it is provided by at least 80% of the hospital trusts in a particular year. This has the effect of reducing the total number of elective conditions against which hospital trusts are compared across the 9 years from

Table 5 **Cost model parameter estimates - using capped costs**

	Costs, cap $K = 10$		Costs, cap $K = 5$	
	Elective	Emergency	Elective	Emergency
Between effects				
Elect. vol. (focal SL)	-0.055^{***}	0.019^{***}	-0.054^{***}	0.017^{***}
	(0.009)	(0.004)	(0.009)	(0.004)
Emerg. vol. (focal SL)	0.007	-0.121^{***}	0.010	-0.119^{***}
	(0.013)	(0.010)	(0.013)	(0.010)
Elect. vol. (other SLs)	0.041	0.134^{***}	0.038	0.126^{***}
	(0.030)	(0.028)	(0.029)	(0.026)
Emerg. vol. (other SLs)	-0.007	-0.098^{**}	-0.006	-0.089^{**}
	(0.034)	(0.031)	(0.033)	(0.029)
Model fit				
Observations	15,339	15,354	15,339	15,354
Marginal R^2	0.096	0.180	0.096	0.181
Conditional R^2	0.516	0.627	0.520	0.634
Bayesian inf. crit.	-71.5	-11,816.2	-1,209.5	-13,498.1

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; Coefficient estimates and standard errors reported for the between effects; Marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

Table 6 **Model parameter estimates - calculated on a set of common HRGs**

	Costs		Length of stay	
	Elective	Emergency	Elective	Emergency
Between effects				
Elect. vol. (focal SL)	-0.031^{***}	0.038^{***}	-0.020^{***}	0.012^{***}
	(0.009)	(0.004)	(0.004)	(0.003)
Emerg. vol. (focal SL)	-0.001	-0.054^{***}	0.022^{***}	-0.055^{***}
	(0.015)	(0.010)	(0.006)	(0.007)
Elect. vol. (other SLs)	0.037	0.111^{***}	-0.004	0.058^{*}
	(0.031)	(0.029)	(0.013)	(0.026)
Emerg. vol. (other SLs)	0.006	-0.128^{***}	0.014	-0.063^{*}
	(0.036)	(0.032)	(0.015)	(0.029)
Model fit				
Observations	15,335	15,354	15,335	15,354
Marginal R^2	0.080	0.152	0.104	0.094
Conditional R^2	0.488	0.624	0.431	0.731
Bayesian inf. crit.	3,331.4	-10,870.3	-18,356.6	-23,147.1

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; Coefficient estimates and standard errors reported for the between effects; Marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

11,969 to 3,888, and emergency conditions from 11,824 to 6,109. This does, however, represent 74% of the total elective activity and 87% of the total emergency activity, respectively. This suggests that those HRGs that are kept in the sample are those that are higher volume.

The results, using the same estimation method as in the paper, are provided in Table 6 ('between-effects' only). As can be seen, the sign, direction and significance of the estimation on the subset of more common HRGs are all as reported in the paper.

6.4. Minimum service line size

In §6.3 above we consider the possibility that the composition of HRGs used to compare hospital trusts (in particular, the inclusion of rarer conditions) may affect the results. Another possibility – which we saw some evidence of in §4 of this document when discussing potential non-linear volume effects – is that the inclusion of hospital trusts which treat only a low volume of activity within

Table 7 **Model parameter estimates - excluding hospital-years with low service line volume**

	Costs		Length of stay	
	Elective	Emergency	Elective	Emergency
Between effects				
Elect. vol. (focal SL)	-0.054^{***}	0.041^{***}	-0.030^{***}	0.019^{***}
	(0.011)	(0.006)	(0.005)	(0.004)
Emerg. vol. (focal SL)	0.007	-0.150^{***}	0.017^{*}	-0.094^{***}
	(0.015)	(0.011)	(0.007)	(0.008)
Elect. vol. (other SLs)	0.037	0.138^{***}	0.006	0.082^{**}
	(0.030)	(0.027)	(0.014)	(0.025)
Emerg. vol. (other SLs)	-0.001	-0.096^{**}	0.013	-0.066^{*}
	(0.034)	(0.031)	(0.016)	(0.028)
Model fit				
Observations	14,151	14,151	14,151	14,151
Marginal R^2	0.088	0.177	0.100	0.110
Conditional R^2	0.528	0.636	0.512	0.766
Bayesian inf. crit.	-1,717.0	-11,263.4	-22,294.0	-23,327.6

$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$; Coefficient estimates and standard errors reported for the between effects; Marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

a particular service line (i.e. which provide no or only a limited scope of service) may be outliers (and not a fair comparison) and may be influencing the results.

To examine this, we have re-run the cost models from the paper on a subset of the data such that only those years in which a trust treats at least 25% of the median elective volume and emergency volume of activity within a particular service-line are included in the sample. This reduces the sample by ∼8% from 15,354 observation to 14,151. The results – provided in Table 7 – are almost identical to those in the paper, suggesting the findings are not heavily influenced by the presence of trust–service-lines with a low volume of activity.

6.5. Single hospital trusts

The analysis in the paper was run on the set of all trusts operating in England. As mentioned in §4 of the paper, trusts may operate multiple hospitals across multiple sites. While often there is often a main hospital site that treats the vast majority of the patients, there are a number of hospital trusts (e.g. Guy's and St. Thomas' in London) where the same trust operates multiple large hospitals. As it is not possible in our data to distinguish between which patients were treated at which site, it could in these cases that specialties and/or elective or emergency patients are split over multiple sites and that the scale effects we identify are affected by this.

To investigate this further, we have repeated the analysis from the paper using a subset of the data corresponding to those trusts that only operate a single hospital site. This has the effect of reducing the sample by 54.0%, from 15,354 observations to 8,285. The results, reported in Table 8, show that even when restricting the sample there is little change in the sign or scale of the main results reported in the paper. As such, in the paper we report the results from all trusts.

Table 8 **Model parameter estimates - subset of trusts operating one hospital site**

	Costs		Length of stay	
	Elective	Emergency	Elective	Emergency
Between effects				
Elect. vol. (focal SL)	-0.051^{***}	0.013^{**}	-0.018^{**}	0.003
	(0.013)	(0.005)	(0.006)	(0.004)
Emerg. vol. (focal SL)	0.039^{*}	-0.126^{***}	0.025^{**}	-0.091^{***}
	(0.020)	(0.015)	(0.009)	(0.011)
Elect. vol. (other SLs)	-0.005	0.134^{**}	-0.017	0.116^{**}
	(0.043)	(0.043)	(0.018)	(0.041)
Emerg. vol. (other SLs)	0.010	$-0.090^{\dagger}$	0.024	-0.117^{*}
	(0.049)	(0.047)	(0.020)	(0.046)
Model fit				
Observations	8,271	8,285	8,271	8,285
Marginal R^2	0.087	0.171	0.114	0.113
Conditional R^2	0.481	0.615	0.405	0.731
Bayesian inf. crit.	1,458.3	-5,259.2	-8,852.9	-11,214.0

$^{\dagger}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; Coefficient estimates and standard errors reported for the between effects; Marginal R^2 describes the proportion of variance explained by the non-random factors only, and conditional R^2 describes the proportion of variance explained by both the non-random and random factors.

References

Baltagi B, Wu P (1999) Unequally spaced panel data regressions with AR(1) disturbances. *Econometric Theory* 15(06):814–823.