COMPETITION COMMISSION

PRIVATE HEALTHCARE MARKET INVESTIGATION

Price-concentration analysis for self-pay patients

Introduction

1. As part of our private healthcare market inquiry, the Annotated Issues Statement (AIS)[1] described the work we have been conducting in relation to each Theory of Harm (ToH). ToH1 in particular posited that there may be adverse effects on competition as a result of hospitals having market power in local areas. This working paper is an addition to the material presented regarding ToH1 in the AIS and in particular the price-concentration analysis (PCA).[2] Its purpose is to provide more detailed explanation of the PCA, and to present updated results and extensions to the initial work presented in the AIS.

2. As set out in the AIS, our initial assessment of ToH1 indicated that certain hospitals did enjoy market power in certain local areas, and that this market power conferred the ability to levy higher prices on self-pay patients for inpatient services. This remains our view. Our analysis estimates an *increase* in a hospital's LOCI measure of around 0.2—which corresponds to a 20 percentage point decrease in average market shares—is likely to result in a price *decrease* of around 3.6 per cent on average. The estimates are statistically significant and robust to a number of modifications that we have considered.

3. This paper is structured as follows. First, the PCA methodology is outlined. Second, the data we use to conduct the analysis is described. The third and fourth sections of the paper discuss the results of our analysis and an assessment of the robustness of those results. The paper then summarizes our conclusions from this

---

[1] See www.competition-commission.org.uk/assets/competitioncommission/docs/2012/private-healthcare-market-investigation/120622_issues_statement.pdf.
[2] See AIS, Appendix B, Annex 3.

work. Appendix 1 provides more details of our processing of the data provided to us by parties.

Methodology

4. Our interest in undertaking this piece of work is to evaluate the relationship between prices and market concentration in local markets.[3] An understanding of this will help explain price differences between hospitals that are otherwise comparable but for the concentration in their local market. It may also allow for the prediction of price responses to changes in local market concentration. Prices and concentration are typically expected to be related such that higher prices prevail in more concentrated markets; however, in any particular instance, there may be countervailing market features that offset the relationship. As a result the nature of the relationship is usually an empirical question and PCA is a well-established methodology for estimating the relationship between these variables.

5. Estimating the price-concentration relationship involves making comparisons between the prices charged by different hospitals and the local market concentration faced. In a simple hypothetical setting, this might be achieved by finding hospitals that are comparable in all respects except for the level of local market concentration faced. Any price difference between such hospitals might then be attributed to a price-concentration relationship. In practice, hospitals and the local markets they operate in are not all exactly comparable and differ in many dimensions which may affect prices charged. This can make simple price comparisons (that neglect these other differences) misleading. PCA addresses this issue by using regression analysis to estimate the price-concentration relationship while controlling for the

[3] References to 'prices' for the remainder of this note should be taken to mean the prices paid by self-pay patients for inpatient hospital services excluding consultant fees and ancillary items.

differences between hospitals and local markets. In effect, the price-concentration relationship is estimated while other factors are 'held fixed'.

6. The particular price-concentration relationship that we seek to understand is that which prevails in the market for private hospital services as a whole. Any nuances to the relationship in certain local areas and/or for certain treatments are not the primary focus here. The general approach to the PCA therefore seeks to characterize a broad relationship that holds across the market rather than the many more micro-relationships that may operate in particular submarkets. In a later section of this paper that deals with the robustness of our results we assess whether the broad relationship is a reasonable generalization of any such micro-relationships for specific treatments.

7. We have taken a 'reduced-form' approach to the PCA.[4] We estimate the following equation:

(Equation 1) $$\ln(price_i) = \beta.concentration_i + \gamma.X_i + u_i$$

8. In this equation, $price_i$ is the price paid for private hospital services by patient *i,* and $concentration_i$ is a measure of local market concentration faced by the hospital that patient *i* visited.[5] The term X_i contains other measurable factors that are specific to patient *i*'s hospital visit and expected to affect the price paid by patient *i*.[6] Factors contained in X_i are referred to as the 'control variables', while $concentration_i$ and X_i collectively are referred to as the 'covariates'. The term u_i represents all 'unobserved' factors that affect prices but that are not included in X_i. The two terms

[4] By reduced form we refer to an approach that does not rely on a particular underlying economic model that is assumed to hold.
[5] The concentration variable therefore varies by hospital site but does not vary between patients that visit the same hospital site.
[6] X_i is a vector that contains several variables.

β and γ represent the 'parameters' that characterize the relationship of each covariate with price.

9. Data on patients can be used to estimate the parameters of Equation 1. In order to do this it is necessary to make certain assumptions. The two key assumptions made are:

Assumption 1: the equation is a reasonable approximation of the relationship between prices and the covariates; and

Assumption 2: the covariates are exogenous (or equivalently, that the covariates are uncorrelated with the unobserved term, u_i).

10. The first assumption relates to particular form of Equation 1, which links the natural logarithm of price to the covariates in a certain way. We use this representation as it produces a model that is simple to interpret and estimate. The natural logarithm allows the analysis to characterize the proportional relationship between prices and concentration through a single parameter (β). This proportional relationship is constant across all treatments that are included in the analysis and thus our attention can focus on a single parameter.

11. The second assumption implies that the covariates, and concentration in particular, are not correlated with any other factors that are not included in the covariates (ie that are included as part of the unobserved term). Further interpretation of Assumption 2 is given later in this working paper, when the robustness of our results to changes in these two assumptions is discussed.

12. If these assumptions hold, then the parameter β can be interpreted as the causal effect of concentration on price—that is, it informs how price may change in response to a change in concentration. More precisely β represents the

(approximate) average percentage change in price following a one unit change in concentration.[7] Under the assumptions stated above, ordinary least squares (OLS) can be used to estimate the parameters (β, γ) in Equation 1. To employ this method, data is required on the prices, concentration and control variables for each patient visit.

Data

13. This section describes the data that has been used in the regression analysis. Three issues are discussed:

(a) the price and concentration variables;

(b) the control variables; and,

(c) the treatments used in the analysis.

Price and concentration variables

14. Two main sources of data have been used in the analysis. Data provided by hospital groups has been used to create the price variable and data provided by Healthcode (an intermediary between hospitals and insurers) has been used to create the concentration variables. Below is a brief description of each dataset and the price and concentration variables. The appendix to this working paper provides further details of the data processing.

15. We received invoice data from the five main hospital groups relating to self-pay patients.[8] We have cleaned and consolidated this data to produce a single dataset of self-pay patient episodes ('the hospital database').[9] An episode is defined as a

[7] The effect is only approximately equal to the percentage change due to the properties of the natural logarithm function.
[8] The five big hospital groups are BMI, HCA, Nuffield, Ramsay and Spire.
[9] See the appendix for more details of the data cleaning. Note that minor changes were made to the hospital dataset since the initial analysis presented in the AIS, and this has had the effect of increasing the available observations in the data.

single visit to hospital. The hospital dataset covers the period 2006 to 2012,[10] and includes information on inpatient episodes at 147 hospital sites.

16. The data we have received from Healthcode is also invoice data, but relates to insured patients and includes records for the majority of hospital groups (ie not just the five large hospital groups). In a similar way to the hospital database, we have cleaned and processed the Healthcode data to produce a single dataset on insured patient episodes covering the period 2006 to 2012.[11] [12] This dataset includes information on inpatient episodes at 173 hospital sites.

17. The price variable used in the regression analysis is the episode price calculated using the hospital dataset. An episode price is defined as the price paid by a self-pay patient for hospital services, excluding the cost of consultant fees and ancillary items.[13]

18. The concentration variables used in the regression analysis have been derived from the Healthcode database and as such are based on insured patients. The Healthcode dataset has been used for this purpose as it represents the most consistent and complete picture of patient journeys that is available to us.[14] [15]Two concentration measures have been used for the purposes of the PCA: the LOCI by patient share and fascia count. The LOCI measure is equal to one minus the weighted average market share of a hospital.[16] We focus throughout this paper on

---

[10] We use data for 2009 to 2012 in the PCA analysis.
[11] See the appendix for more details of the data cleaning.
[12] We use data for 2009 to 2012 in the PCA analysis.
[13] There are minor differences in this definition across the data for each hospital group (eg for BMI data we could not exclude ancillary items) but such differences are expected to be minor.
[14] It provides a more complete picture of patient journeys than the hospital dataset in at least three respects. First, the number of insured patients greatly exceeds the number of self-pay patients. Second, the hospital database contains only information for the five main hospital groups, and therefore omits information for other operators. Third, in order to achieve consistency across the five main party datasets, our data processing led to a higher proportion of episode exclusions from the hospital dataset than the Healthcode dataset.
[15] Concentration measures based on insured patients and self-pay patients are expected to be highly correlated.
[16] Market shares are calculated for each submarket, defined as an outward postcode area, and are then aggregated to a hospital level by a weighted average, with the weights reflecting the importance of each submarket to the hospital. This methodology is described in more detail in the AIS Appendix B, Annex 2.

the LOCI measure calculated with market shares by patient episodes, and accounting for the network ownership of hospital groups. Fascia count has been computed as the count of general private hospital and PPU fascia within three distance bands from the focal hospital: 0–9 miles, 9–17 miles and 17–26 miles.[17] More details on the concentration measures can be found in the AIS Appendix B, Annex 1 and Annex 2.

19. The concentration measures have been constructed once, using the period 2009 to 2012 as a reference period.[18] To match this, only price data from the same period has been used in this analysis. We use only inpatient episodes for the concentration measures and the PCA.

Control variables

20. Equation 1 specified a group of control variables, X_i. This group of variables should include the factors that are expected to affect prices, as well as being correlated with the concentration measures. If factors that meet these conditions are not included in the variables, Assumption 2 is less likely to hold. Factors that affect supply and demand conditions for private healthcare services are typical candidates for control variables. We have considered the following control variables:

(a) year dummies, to account for any movement in the average price over time;

(b) operator dummies, to account for any differences in the average price between the five large hospital groups;[19]

(c) treatment dummies, to account for differences in average price between the different treatments included in the analysis;

---

[17] The Healthcode dataset itself is not required to calculate the fascia counts, since only the location of hospitals is required and this is publically available. However, Healthcode dataset was used to establish the median catchment area for UK hospitals (17 miles).

[18] In the AIS it was stated that 2009-2011 was used as the reference period. This was incorrect, and 2009-2012 data was in fact used.

[19] Such price differences may arise as a result of different prices charged and also from the minor differences in the datasets (eg due to recording practices or minor inconsistencies in our consolidated hospital dataset).

(d) patient age, patient gender and the number of nights per episode, to account for differences in the individual circumstances of each patient;[20]

(e) average direct cost of the hospital (logged), to account for differences in input or labour costs;[21] and

(f) location dummies, to account for any differences in supply and/or demand specific to local areas.[22]

21. The data for variables *(a)* to *(d)* above comes directly from the hospital dataset. The data for the cost variable *(e)* has been submitted to us by the five large hospital operators in response to the Financial Questionnaire and we have cleaned and matched this data to the hospital dataset. The location variables, *(f)*, have been created by linking the postcode of each treating hospital to the appropriate geographic classification. This linking was done using data provided by the Office for National Statistics.

22. Different levels of geographic classification are available for the location dummies. We have investigated using different classifications but focus in this note on 'NUTS2', which is a classification developed by the European Union and classifies the hospitals in our hospital dataset into 34 separate areas. The robustness of our analysis to this choice of classification is considered later.

Treatments

23. The hospital dataset contains patient episodes that relate to many different treatments (eg hip operation, cataract surgery etc). Each treatment is defined by its 'CCSD code', a five-digit code that has a corresponding description. Treatments

---

[20] The number of nights per episode may proxy for the severity of a particular treatment. For example, patients receiving hospital services relating to a hip replacement may stay a larger number of nights if the treatment is more complex or severe.
[21] This is calculated as the total direct cost of each hospital site, divided by the total number of patients (itself the sum of inpatient, day patient and outpatient visits). Cost data was available for almost all hospitals in our analysis. For hospitals with missing cost data, we have imputed the data on the basis of hospitals owned by the same operator in the region.
[22] Differences specific to each area might include demand and supply conditions such as population, demographics, and the supply of NHS services.

could in principle be analysed individually to assess the price-concentration relationship for each but with several hundred treatments this is not practical. We have therefore considered a small group of important treatments that we consider likely to reflect the overarching price-concentration relationship across the market as a whole. These treatments are grouped together in the analysis, and our understanding of the market suggests that there is a reasonable degree of supply-side substitution across these treatments.[23]

24. We have selected eight 'focal treatments'. These were selected as treatments with a high number of inpatient visits in our hospital dataset.[24] The focal treatments account for around half of the patient episodes and revenue in the hospital dataset, and as such may reflect the important features of the price-concentration relationship across the market as a whole. Our data provides information on prices charged for each of these focal treatments for at least four of five main parties and between 64 and 125 hospital sites (depending on the treatment). The robustness of our results to the choice of these particular treatments is considered later.

25. The focal treatments are listed in Table 1 together with summary statistics in Table 2.

[23] For more details on product markets see the AIS, Appendix A.
[24] The focal treatments are the top eight treatments over the period 2006 to 2012, and eight of the top nine treatments over the period 2009 to 2012. Over this shorter period, gastric bypass (G3100) ranks eighth (one above cataract surgery, C7122) but was only provided by 39 hospitals as compared with the focal treatments, each of which was provided by at least 76 hospitals.

TABLE 1 **Focal treatments**

CCSD code	*Abbreviated description*	*Speciality*	*Observations*	*Revenue £*	*Hospital sites with invoices*
C7122	Cataract surgery	Ophthalmology	1,137	1,480,587	107
E0260	Rhinoplasty following trauma	Plastic surgery	2,376	4,913,131	109
G3080	Gastric banding	General surgery	2,950	12,722,695	76
J1830	Removal of gallbladder	General surgery	1,412	4,987,644	130
M6530	Prostate resection	Urology	1,808	6,768,967	130
T2000	Inguinal hernia surgery	General surgery	2,070	3,061,330	137
W3712	Hip replacement	Trauma and orthopaedics	5,834	49,552,616	138
W4210	Knee replacement	Trauma and orthopaedics	3,250	29,898,737	130
All focal treatments			20,837	113,385,707	145
All treatments			46,681	207,457,785	147

Source: CC analysis.

Note: Numbers may not sum due to rounding. Figures taken from the cleaned hospital dataset (2009–2012).

TABLE 2 **Descriptive statistics for focal treatments**

CCSD code	*Abbreviated description*	*Average price £*	*Median price £*	*Min price £*	*Max price £*	*Std deviation*
C7122	Cataract surgery	1,302	1,238	690	3,057	355
E0260	Rhinoplasty following trauma	2,068	1,730	950	6,525	910
G3080	Gastric banding	4,313	4,675	1,455	7,600	1,383
J1830	Removal of gallbladder	3,532	3,513	1,890	5,917	578
M6530	Prostate resection	3,744	3,694	2,075	6,500	493
T2000	Inguinal hernia surgery	1,479	1,458	899	2,161	219
W3712	Hip replacement	8,494	8,399	5,818	11,989	996
W4210	Knee replacement	9,200	9,153	5,484	15,215	1,155
All focal treatments		5,442	4,851	690	15,215	3,192

Source: CC analysis.

Note: Numbers may not sum due to rounding. Figures taken from the cleaned hospital dataset (2009–2012).

Results

26. This section sets out the results of estimating Equation 1 under Assumptions 1 and 2. We use the data described in the previous section on focal treatments. Estimation results of specifications that use LOCI as the concentration measure are considered first, followed by specifications that use the fascia count variables as the concentration measure. For both sets of specifications we consider different choices of control variables.

LOCI

27. Table 3 below sets out the results of the regressions using LOCI as the concentration measure. Specification (1) includes year, operator and treatment dummies as control variables. Specification (2) and (3) use additional control variables:

specification (2) adds in the patient-level controls (age, gender, nights) and specification (3) then adds in the cost variable and the regional dummies (not shown in the table).

TABLE 3 **Regression results, LOCI**

	(1)		*(2)*		*(3)*	
	Coefficient	*Std error*	*Coefficient*	*Std error*	*Coefficient*	*Std error*
LOCI	–0.162*	0.068	–0.163*	0.068	–0.180***	0.053
Year dummy: =1 if 2010	–0.005	0.012	–0.005	0.012	–0.001	0.01
Year dummy: =1 if 2011	0.031**	0.011	0.033**	0.011	0.040***	0.01
Year dummy: =1 if 2012	0.043**	0.013	0.045**	0.014	0.052***	0.012
Operator dummy: =1 if HCA	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Operator dummy: =1 if Nuffield	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Operator dummy: =1 if Ramsay	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Operator dummy: =1 if Spire	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Treatment dummy: =1 if Cataract surgery	–1.902***	0.039	–1.888***	0.042	–1.888***	0.038
Treatment dummy: =1 if Rhinoplasty following trauma	–1.460***	0.059	–1.429***	0.061	–1.431***	0.061
Treatment dummy: =1 if Gastric banding	–0.725***	0.048	–0.698***	0.047	–0.706***	0.05
Treatment dummy: =1 if Removal of gallbladder	–0.878***	0.014	–0.858***	0.016	–0.867***	0.016
Treatment dummy: =1 if Prostate resection	–0.816***	0.014	–0.813***	0.014	–0.820***	0.015
Treatment dummy: =1 if Inguinal hernia surgery	–1.751***	0.015	–1.739***	0.017	–1.740***	0.018
Treatment dummy: =1 if Knee replacement	0.081***	0.01	0.080***	0.01	0.078***	0.01
Patient sex			–0.008	0.005	–0.008	0.004
Patient age			0.000	0.000	0.000	0.000
Episode number of patient nights			0.004	0.002	0.005*	0.002
ln(average direct cost)					–0.016	0.032
[Location dummies]					[Not shown]	[Not shown]
Constant	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
R-squared	0.91		0.91		0.917	
N	20720		20720		20720	

Source: CC analysis.

Note: Numbers may not sum due to rounding. Base categories for dummy variables are BMI, 2009 and hip replacement. Standard errors are clustered by hospital site. Blank entries indicate that the covariate is not included in the specification. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

28. Looking at specification (1), coefficient on the LOCI variable, –0.162, reflects the estimated price-concentration relationship—this is the estimate of β from Equation 1. It indicates the likely impact on prices of a change in LOCI of one unit. The coefficient on the LOCI variable can therefore be interpreted as the average impact on prices for a change in LOCI from zero (monopoly) to one (perfect competition).

Specification (1) indicates that this would cause a reduction in prices of around 16 per cent. The effect is statistically significant at the 5 per cent level.

29. Looking now at specification (2) and (3), in a similar manner to specification (1), these two specifications estimate a statistically significant price-concentration relationship, with prices expected to fall by around 16 to 18 per cent following an increase in the LOCI from zero to one. The estimate in specification (3) is statistically significant at the 0.1 per cent level.

30. Changes of LOCI of one unit, as reported by the regressions, are purely an artefact of the regression methodology. While the difference in market structure between monopoly and perfect competition is a useful benchmark, such a comparison or change in market structure is extreme and unlikely to ever occur in practice; moreover, there are no hospitals in our dataset with a LOCI of zero or a LOCI of one. When interpreting the results it is therefore important to consider the likely impact on prices of more modest changes in LOCI—this can be achieved by scaling the estimated effect linearly. For example, according to estimates of specification (3) an increase in the LOCI of 0.5 (ie a 50 per cent decrease in the weighted average market share) will cause price reductions of around 9 per cent and an increase in LOCI of 0.2 (ie a 20 per cent decrease in the weighted average market share) is estimated to cause a price reduction of around 3.6 per cent.

Fascia count

31. Table 4 below sets out the results of the specifications that use the fascia count variables as the concentration measure. Specification (4) includes year, operator and treatment dummies as control variables. Specification (5) and (6) use additional control variables: specification (5) adds in the patient-level controls (age, gender,

nights) and specification (6) then adds in the cost variable and the regional dummies (not shown in the table).

TABLE 4 **Regression results, fascia count**

	(4)		*(5)*		*(6)*	
	Coefficient	*Std error*	*Coefficient*	*Std error*	*Coefficient*	*Std error*
Fascia count (0–9 miles)	–0.012	0.009	–0.013	0.009	–0.004	0.008
Fascia count (9–17 miles)	–0.003	0.003	–0.003	0.003	–0.002	0.003
Fascia count (17–26 miles)	0.002	0.002	0.002	0.002	0.001	0.003
Year dummy: =1 if 2010	–0.007	0.012	–0.006	0.012	–0.002	0.011
Year dummy: =1 if 2011	0.029*	0.011	0.031**	0.012	0.038***	0.01
Year dummy: =1 if 2012	0.042**	0.014	0.045**	0.014	0.051***	0.012
Operator dummy: =1 if HCA	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Operator dummy: =1 if Nuffield	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Operator dummy: =1 if Ramsay	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Operator dummy: =1 if Spire	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
Treatment dummy: =1 if Cataract surgery	–1.905***	0.039	–1.890***	0.041	–1.889***	0.037
Treatment dummy: =1 if Rhinoplasty following trauma	–1.463***	0.058	–1.432***	0.061	–1.434***	0.061
Treatment dummy: =1 if Gastric banding	–0.736***	0.053	–0.708***	0.052	–0.709***	0.051
Treatment dummy: =1 if Removal of gallbladder	–0.878***	0.014	–0.857***	0.016	–0.863***	0.016
Treatment dummy: =1 if Prostate resection	–0.815***	0.013	–0.811***	0.013	–0.815***	0.015
Treatment dummy: =1 if Inguinal hernia surgery	–1.750***	0.015	–1.737***	0.017	–1.737***	0.018
Treatment dummy: =1 if Knee replacement	0.080***	0.01	0.079***	0.01	0.077***	0.01
Patient sex			–0.007	0.005	–0.007	0.004
Patient age			0.000	0.000	0.000	0.000
Episode number of patient nights			0.004	0.002	0.005*	0.002
ln(average direct cost)					–0.023	0.032
[Location dummies]					[Not shown]	[Not shown]
Constant	[✂]	[✂]	[✂]	[✂]	[✂]	[✂]
R-squared	0.909		0.909		0.916	
N	20837		20837		20837	

Source: CC analysis.

Note: Numbers may not sum due to rounding. Base categories for dummy variables are BMI, 2009 and hip replacement. Standard errors are clustered by hospital site. Blank entries indicate that the covariate is not included in the specification. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

32. The regression using fascia count as a measure of concentration can be interpreted in a similar way to the LOCI regressions, except that there are now three concentration variables, and each associated coefficient reflect a change in fascia count of one unit. The model therefore reports the average impact on prices in response to changes in the number of competing fascia at three different distance bands. A one unit change in fascia count implies one additional rival hospital in the

relevant distance band (ie a much less extreme change in concentration than from monopoly to perfect competition as in the LOCI specifications).

33. Specification (4) indicates that an additional fascia within a 9-mile radius is expected to reduce prices by around 1.2 per cent, but more muted price effects of less than 1 per cent from additional fascia at farther distances (indicated by the coefficients on the fascia count variables corresponding to 9–17 miles and 17–26 miles). Specifications (5) and (6), which add more covariates to specification (4), report similar findings. None of the specifications using the fascia count estimate statistically significant price-concentration relationships at the 5 per cent level, an issue that is discussed shortly.

34. Comparing the estimated price-concentration relationships in the LOCI specifications and fascia count specifications reveals two main differences. First, the estimated relationships are statistically significant in the LOCI specifications while they are not in the fascia count specifications. Second, the estimated relationships in the LOCI specifications are larger in magnitude. To see this latter point, increases in the LOCI of between 0.2 and 0.5 can roughly be equated to an increase of one fascia (eg from no rival fascia to one rival fascia, or one rival fascia to two rival fascia) and the predicted price responses are between 3.6 and 9 per cent as compared with between 0 and 1.3 per cent, respectively. In other words, the predicted price responses are almost three times greater or more according to the LOCI specifications.

35. There are several potential explanations for these two differences. One explanation is that LOCI is a more accurate measure of market concentration (and thus a proxy to market power) than fascia count. This is in line with our reasoning for using LOCI in this inquiry—it differentiates between the strength of competitors (fascia count

does not) and it does not rely on fixed distance bands (fascia count does).[25] With these points in mind, while the fascia count is an indicator of local market concentration, it is a less refined measure than the LOCI in that it cannot detect more nuanced differences in market concentration. Intuitively, since the measured price responses to changes in concentration may be relatively moderate (as indicated by the LOCI specifications), delineating this relationship using the less refined fascia count measure is likely to prove challenging. This intuitive argument for the differences in results is also supported by our interpretation of the regressions. In particular, it is expected that estimated price responses to a less accurately measured variable (ie a more noisy signal of the underlying variable) will be more muted and less statistically significant than the estimated price responses to a more accurately measured variable (ie a less noisy signal of the underlying variable).[26] We therefore view the LOCI specifications, as compared with the fascia count specifications, as providing a better reflection of the price-concentration relationship.

36. The remainder of this working paper therefore assesses the robustness the LOCI specifications to various modifications. We focus on specification (3) which incorporated all of our control variables. Based on this specification, the estimated price reduction is around 18 per cent for a change in LOCI from zero (monopoly) to one (perfect competition). This specification is referred to as the 'baseline specification'.

Robustness of the results

37. This section considers how estimates from the baseline specification may be affected by various modifications to the methodology or data. The purpose of making these

[25] For further comparisons between LOCI and other concentration measures see the AIS, Appendix B, Annex 2.
[26] In econometric terms this is sometimes referred to as 'classical measurement error' in a covariate.

modifications is to assess how sensitive our estimates are, and thus whether the baseline specification provides a reasonable characterization of the price-concentration relationship that we seek to understand. Four issues are addressed:

(a) the functional form of the model (Assumption 1);

(b) exogeneity of the covariates (Assumption 2);

(c) the calculation of LOCI in relation to missing invoices; and

(d) other modifications relating to the data.

38. All analysis focuses on changes to the baseline specification (specification (3) in Table 3).

Functional form of the model

39. Assumption 1 in the methodology section of this paper was that Equation 1 was a reasonable approximation to the relationship between price and the covariates. The particular representation used, pooling all treatments together and using the natural logarithm, was chosen on the basis that it was able to represent the price-concentration relationship in a simple manner. Alternative specifications have been considered to assess whether the baseline specification in the form chosen adequately represents the price-concentration relationship. The following alternative specifications have been considered:

(a) a specification that allows for different relationships for each focal treatment;[27] and

(b) a specification that does not impose the natural logarithm transformation and instead uses a linear specification (and also allows for different relationships for each focal treatment).[28]

---

[27] That is, the model is estimated for each of the focal treatments separately.

[28] This model has only been estimated separately for each treatment. Estimating Equation 1 for all focal treatments together but without the natural logarithm would produce a model that imposes the same 'level impact' (rather than percentage impact) on

40. The results of these different specifications are reported in Tables 5 and 6 below.

TABLE 5 **Regression results, LOCI—treatment-by-treatment analysis**

	(3)	*C7122*	*E0260*	*G3080*	*J1830*	*M6530*	*T2000*	*W3712*	*W4210*
LOCI	–0.180***	–0.006	–0.352**	–0.394**	–0.147*	–0.042	0.025	–0.097	–0.133*
[Other covariates not shown]									
R-squared	0.917	0.549	0.587	0.418	0.336	0.347	0.34	0.291	0.256
N	20720	1137	2370	2947	1407	1808	2060	5758	3233

Source: CC analysis.

Note: Numbers may not sum due to rounding. Dependent variable and covariates for each specification are the same as specification (3). Standard errors are clustered by hospital site in all specifications. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

TABLE 6 **Regression results, LOCI—treatment-by-treatment analysis with no log transformation**

	(3)	*C7122*	*E0260*	*G3080*	*J1830*	*M6530*	*T2000*	*W3712*	*W4210*
LOCI	–0.180***	41.5	–856.0***	–1104.3*	–540.7*	–147.8	26.2	–900.3*	–1275.4**
[Other covariates not shown]									
R-squared	0.917	0.509	0.609	0.331	0.336	0.349	0.32	0.284	0.268
N	20,720	1,137	2,370	2,947	1,407	1,808	2,060	5,758	3,233
Average episode price		1302	2068	4314	3531	3744	1479	8484	9196
LOCI marginal effect as percentage of average episode price (%)		3.2	–41.4	–25.6	–15.3	–4.0	1.8	–10.6	–13.9

Source: CC analysis.

Note: Numbers may not sum due to rounding. For the treatment-level specifications the dependent variable is the level of episode price and the covariates are the same as the specification (3) with the exception of the cost variable which is not logged. Standard errors are clustered by hospital site in all specifications. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

41. Table 5 shows results of the baseline specification but estimated separately on the data for each treatment. There is some variation in the estimated price-concentration relationship by treatment, which varies from 0.025 to –0.394. The statistically significant estimates at the 5 per cent level range from –0.133 to –0.394 (ie the positive estimate is statistically insignificant), and the average of these is –

prices for changes in the LOCI (eg £100 price reduction as a consequence of a unit change in LOCI). This is seen as less likely to adequately characterize the price-concentration relationship because of the large differences in average price for each treatment.

0.241. These estimates are comparable in magnitude to the baseline specification estimate of –0.180.

42. Table 6 also shows estimates that consider each treatment in isolation, but in these specifications the dependent variable is the level of episode price (ie without the natural logarithm); all covariates remain the same as the baseline specification except for cost which also has had the log transformation removed. Table 6 shows in the final row the estimated price-concentration relationships for these specifications in percentage terms so that they are comparable to the estimates in Table 5 (ie these percentages can be compared with –18.0 per cent, which corresponds to baseline specification estimate of –0.180). In a similar way to Table 5, the results of Table 6 indicate some variation by treatment, ranging from 3.2 to –41.4 per cent. The statistically significant effects at the 5 per cent level range from –10.6 to –41.4 per cent (as with Table 5, the positive estimates are statistically insignificant), and the average of these is–21.4 per cent. These estimates are comparable in magnitude to the estimates from the baseline specification of –18.0 per cent.

43. Taken together, the results of Tables 5 and 6 indicate that the baseline specification characterizes the average price-concentration relationship relatively well. While the differences at the treatment level highlight that the baseline specification is (as expected) a simplification of several micro-relationships, it does not suggest that this is unreasonable.

Exogeneity of the covariates

44. Assumption 2 of the model is that the covariates, and LOCI in particular, are exogenous. In other words, the covariates are uncorrelated with other factors that are unobserved. If this assumption does not hold, one or more covariates is said to

be endogenous. This might happen if there are factors directly affecting prices that are also correlated with concentration but not included in the covariates ('omitted variables'). Depending on the nature of the endogeneity—the cause, the interrelationship between price and the covariates, and the degree of endogeneity—the resulting bias may be upwards, downwards or of a negligible magnitude.

45. In PCA studies it is often considered whether the concentration measure, LOCI in the baseline specification, suffers from endogeneity. This is often motivated by the reasoning given above regarding omitted variables and it is this potential source of endogeneity that we focus on here. For this to cause meaningful bias in the estimated relationship, there would need to be an omitted factor that directly and substantially affects prices and is also correlated with LOCI (either through simple correlation, or because the factor directly affects LOCI as well as price).

46. In the current case, we are of the view that any endogeneity bias is likely to be limited. In other words, LOCI and the other covariates are considered approximately exogenous. Our reasoning relates primarily to the use of regional dummy variables. These will capture any differences in the average market conditions hospitals face in different regions—this will include, for example, differences in population, demographics and other potential competitive constraints not reflected in the LOCI (such as the NHS). In the case of supply-side factors, the cost variable will also go further and capture any price differences that result from within-region differences in hospital costs. Any endogeneity bias is therefore likely to be limited to within-region differences in supply or demand conditions that have a direct and substantial effect on price, and are correlated with LOCI (through the market shares). These within-region factors are thought to be limited, and thus unlikely to induce substantial endogeneity bias.

47. In addition to the reasoning above, we have sought to directly assess the direction and magnitude of any bias that might arise from endogeneity. To this end we have considered an instrumental variables (IV) approach.[29] This requires additional variables, known as instruments, to be used in the regression. For the IV approach and associated instruments to adequately correct and test for endogeneity, the instruments must satisfy a number of conditions. In general, finding variables that meet these conditions can be challenging and if the instruments do not satisfy the conditions the IV technique does not guarantee improvements to the specification. The three conditions required of instruments are:

(a) the instruments should be correlated with the potentially endogenous variable (LOCI in the baseline specification)—instruments that meet this condition are said to be 'relevant';[30]

(b) the instruments should be uncorrelated with the unobserved term in Equation 1—instruments that meet the second condition are said to be 'exogenous'; and

(c) the instruments should themselves be excluded from the covariates in the price equation—instruments that meet this condition are said to be 'excluded'.

48. We have considered the following two instruments: the distance to the nearest rival hospital; and, the distance to the nearest hospital under common ownership. These are variables that are likely to be 'relevant'—and thus satisfy condition *(a)* above—since hospitals that are farther away from rival hospitals and/or closer to hospitals under common ownership are likely to have higher market shares and lower LOCI (producing a correlation between the instruments and LOCI). This is directly evident from the LOCI methodology. It is therefore a question of whether the conditions *(b)* and *(c)* above hold.

---

[29] Later in this paper we also report estimates using a specification with more disaggregated regional dummies. To an extent this specification will also address endogeneity concerns since it effectively controls for differences between smaller regions, which is likely to mitigate further any within-region differences. The results of that specification are consistent with the IV regressions presented here.

[30] To be precise, this correlation should be conditional on the exogenous covariates.

49. Condition *(b)*, that these distance variables are exogenous, requires the variables to be uncorrelated with any of the presumed causes of endogeneity. As argued above, because of the inclusion of location dummies, there are not thought to be substantive omitted variables. However, it might be argued that there are within-region differences in demand within each region that substantially affect prices charged. If these within-region differences were also to affect LOCI (through the market shares), endogeneity bias may arise. Taking this argument, the distance instruments would satisfy condition *(b)* if they were uncorrelated with differences in within-region demand conditions. While the distances between (any) hospital sites may not satisfy this requirement—for example, because there are more hospitals that are closely located in areas of high demand—the relative location of rival hospitals and/or hospitals that are under common ownership may satisfy this requirement if past mergers and acquisitions are unrelated to within-region differences in demand. This is assumed to hold approximately for the purposes of this analysis, but is addressed statistically later.

50. Condition *(c)* will hold if the distance variables themselves are not thought to directly affect prices in Equation 1. This would hold if LOCI captured all of the pricing power possessed by a hospital, and the distance measures did not reflect another dimension of market power. As the LOCI measure incorporates geographic relationships between hospitals in its calculation, we think it is reasonable to exclude the distance variables from Equation 1 and thus assume condition *(c)* holds.

51. To the extent that the instruments satisfy all three conditions, it is possible to assess the direction and size of any endogeneity bias. It is also possible to test whether the assumption in question, that LOCI is exogenous (Assumption 2), holds.

52. Table 7 below shows the results of four specifications, (7)–(10), that assume the distance instruments are valid, along with several statistical tests. Specification (7) and (8) use as instruments the distance to the nearest rival hospital or the distance to the nearest hospital under common ownership, respectively; specifications (9) and (10) use both distance instruments in conjunction.

TABLE 7 **Regression results, LOCI—endogeneity analysis**

	(3)	*(7) IV, distance to nearest rival hospital*	*(8) IV, distance to nearest hospital under common ownership*	*(9) IV, both instruments*	*(10) GMM, both instruments*
LOCI	–0.180***	–0.549	–0.241	–0.295*	–0.363*
[Other covariates not shown]					
R-squared	0.917	0.912	0.916	0.916	0.914
N	20,720	20,720	20,720	20,720	20,613
Test of null hypothesis that instruments are irrelevant (F-statistic)		2.894	16.513	19.946	19.833
Test of null hypothesis that the covariates are exogenous (p-value)		0.175	0.657	0.257	1.000
Test of null hypothesis that the instruments are exogenous (p-value)					1.000

Source: CC analysis.

Note: Numbers may not sum due to rounding. IV refers to the two-stage least squares (2SLS) estimator. GMM refers to the two-step efficient GMM estimator; hospital sites with less than 30 observations are excluded when using this estimator. The test of instrument relevance is the F-statistic from the first-stage of 2SLS regressions. The test of covariate exogeneity is Wooldridge's (1995) robust score test for 2SLS models and the C (difference-in-Sargan) statistic for GMM models. The test of instrument exogeneity is the Hansen J-statistic. All specifications and test statistics were computed with Stata's ivregress command. Standard errors are clustered by hospital site in all specifications. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

53. Estimates for specifications (7)–(10) range from –0.241 to –0.549. Two of the four specifications report statistically insignificant estimates at the 5 per cent level. The general increase in the magnitude of the estimated price-concentration relationship, relative to the baseline estimates, indicates (on the basis of the distance instruments) that any bias stemming from endogeneity is likely downwards—ie the baseline regression may understate the true price-concentration relationship. The fact that the estimated relationships are generally less statistically significant is a

likely consequence of the IV methodology, which by construction is less precise than standard OLS techniques.[31]

54. Table 7 also shows the results of three statistical tests: first, that the instruments are relevant (condition *(a)* above); second, that the covariates in the baseline specification are exogenous (Assumption 2); and third, that the instruments are exogenous (condition *(b)* above). Specifications (7) and (8) estimate the baseline specification using the distance to nearest rival hospital or the distance to nearest hospital under common ownership, respectively, as instruments. The results of the test for instrument relevance indicate that the distance to the nearest rival hospital may not be a relevant instrument (relatively low F-statistic of 2.894), but that distance to the nearest hospital under common ownership is a relevant instrument (high F-statistic of 16.513). This suggests that the first instrument (distance to nearest rival hospital) may not be relevant and that there is a stronger case for using the second instrument (distance to nearest hospital under common ownership). These two specifications also cannot reject the hypothesis that the covariates in the baseline specification are exogenous (p-values of 0.175 and 0.657, respectively) suggesting that any endogeneity bias is limited.

55. Specifications (9) and (10) both use the two distance instruments together, but the specifications differ in their estimation method.[32] The first two statistical tests (of relevance, and of covariate exogeneity) for these specifications indicate that the instruments are jointly relevant (F-statistics of around 19 for each) and, as with specifications (7) and (8), that the covariates are exogenous (p-values of 0.257 and 1.000, respectively). One advantage of specifications (9) and (10) that use both instruments together (as compared with the specifications that only use one

[31] The standard errors for the estimated LOCI parameters (not shown in the table) for specifications (7)–(10) are: 0.396, 0.157, 0.120 and 0.180, respectively. The associated p-values are 0.166, 0.124, 0.014 and 0.044, respectively.
[32] Specification (9) uses standard IV techniques known as two-stage least squares. Specification (10) uses two-step GMM.

instrument) is that it is possible to perform the third statistical test—that is, whether the instruments are exogenous (condition *(b)* above). In practice, technical reasons limit the applicability of this test to specification (10).[33] This third test, performed for specification (10), indicates that there is not sufficient statistical evidence to reject that the instruments are exogenous (p-value of 1.000)—ie the instruments are likely to satisfy condition *(b)* described above, and as such may be considered valid.

56. In summary, the baseline specification is not thought to suffer from substantial endogeneity bias. Our reasoning for this primarily relates to the inclusion of location dummies. Estimated relationships from specifications using the IV approach, based on the distance instruments, suggest that any endogeneity bias may mean that the baseline specification understates the magnitude of the true relationship, but not by a large degree. However, statistical tests cannot reject that the covariates in the baseline specification are exogenous (ie indicating insubstantial endogeneity bias), and also suggest that the instruments are valid. This supports the conclusion that any bias arising from endogeneity is limited.

The effect of missing invoices on the LOCI

57. In the AIS, it was noted LOCI may be less well measured in certain areas because some hospitals do not report their invoice information to Healthcode. This means that the Healthcode dataset does not record all patient visits. We have considered the impact of this issue on the estimated price-concentration relationship.

58. The specific concern is the hospitals that do not use Healthcode (the 'missing invoice hospitals') are not represented in the Healthcode invoice data at all. As a consequence, in areas that missing invoices hospitals draw patients from (but we do not observe the invoices), the market share calculations may be subject to some

[33] The usual over-identification test procedures for 2SLS models are not valid with clustered standard errors.

bias. In particular, the market shares for these areas will be overstated for hospitals that do report invoices in Healthcode. In principle there are reasons to expect this bias to be limited since the majority of hospitals do use Healthcode (of the 223 hospitals considered in our inquiry, 173 use Healthcode while 50 do not), and it is typically the smaller hospitals that do not use Healthcode (of the 50 hospitals that do not use Healthcode, 41 are PPUs). However, we attempt to test what impact the missing invoices relating to these 50 hospitals (via the impact on the LOCI) may have on our PCA results.

59. To assess the issue we have first identified those hospitals that are most likely to be affected by this issue (the 'potentially affected hospitals'). Potentially affected hospitals are defined as those that have one or more missing invoice hospitals located within a 17-mile radius. There are 66 potentially affected hospitals, of which the majority only have one or two missing invoice hospitals located in the 17-mile radius. Using this information we have then re-estimated the baseline specification but excluded the potentially affected hospitals from the analysis. A comparison between the estimates of the baseline specification including and excluding the potentially affected hospitals should inform the impact of missing invoices on the PCA results.

60. Table 8 below compares the results of this analysis against the baseline estimates.

TABLE 8 **Regression results, LOCI—removing affected hospitals**

	(3)	*(11) Excluding potentially affected hospitals*
LOCI	–0.180***	–0.275***
[Other covariates not shown]		
R-squared	0.917	0.91
N	20,720	13,192

Source: CC analysis.

Note: Numbers may not sum due to rounding. Standard errors are clustered by hospital site in all specifications. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

61. Specification (11) shows the estimated relationship once the potentially affected hospital sites are removed from the data as compared with specification (3) that includes all hospitals. The effect of removing potentially affected hospitals is to increase the magnitude of the estimated relationship to –0.275 from –0.180. The statistical significance of the estimated relationship remains at the 0.1 per cent level. This indicates that any bias due to the missing invoices is unlikely to distort the baseline estimates to a large degree, and may mean that the baseline regression understates the magnitude of the true price-concentration relationship (ie the estimated price reductions may be greater).

Other modifications relating to the data

62. In this final section we consider a number of other issues which may affect the baseline specification. Four issues are assessed:

(a) the exclusion of irregular episodes from the hospital dataset;[34]

(b) the choice of focal treatments;

(c) the choice of geographic classification for the regional dummies; and

(d) the pooling together of data from all operators.

63. For each of these we have considered an alternative scenario and re-estimated the baseline specification. The scenarios we have considered for each of the above issues are, respectively: reintroducing the excluded irregular episodes (specification (12)); allowing all treatments to enter the model and not just focal treatments (specification (13)); selecting the more granular geographic classification of NUTS3 rather than NUTS2 (specification (14);[35] and, estimating the baseline specification separately for each operator's data. Tables 9 and 10 below compare the results of these scenarios to the baseline specification.

---

[34] The particular exclusions we consider relate to reason *(c)* given in paragraph 8 in the appendix to this paper.
[35] NUTS3 categories the hospitals in our dataset into 80 different regions. This is over double the number of regions as per the NUTS2 classification.

TABLE 9 **Regression results, LOCI—other modifications to the data**

	(3)	*(12) No irregular episode exclusions*	*(13) All treatments*	*(14) NUTS3 regional dummies*
LOCI	–0.180***	–0.267***	–0.091*	–0.204**
[Other covariates not shown]				
R-squared	0.917	0.564	0.895	0.921
N	20,720	23,642	46,390	20,720

Source: CC analysis.

Note: Numbers may not sum due to rounding. Dependent variable and covariates for each specification are the same as specification (3) with the exception of specification (14) which uses NUTS3 regional dummies rather than NUTS2 regional dummies. Standard errors are clustered by hospital site in all specifications. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

TABLE 10 **Regression results, LOCI—operator-by-operator analysis**

	(3)	*BMI*	*HCA*	*Nuffield*	*Ramsay*	*Spire*
LOCI	–0.180***	–0.062	–0.386	–0.470***	0.091	–0.107
[Other covariates not shown]						
R-squared	0.917	0.968	0.925	0.874	0.982	0.964
N	20,720	5,463	362	6,816	1,349	6,730

Source: CC analysis.

Note: Numbers may not sum due to rounding. Dependent variable and covariates for each specification are the same as specification (3). Standard errors are clustered by hospital site in all specifications. ***/**/* indicates statistical significance at the 0.1%/1%/5% level.

64. Table 9 is considered first. The modifications considered in specifications (12), (13) and (14) above show estimated price-concentration relationships of magnitude between –0.091 and –0.267, each of which are statistically significant at levels between 5 and 0.1 per cent.

65. Specification (12) shows that by reincluding data that we had deemed appropriate to exclude—hence increasing the sample size to 23,642—has the effect of increasing the magnitude of the estimated price-concentration relationship, but the model notably fits much worse, with an R^2 value (0.564) that is significantly lower than the baseline regression. This is consistent with our reasoning for making the exclusions; that is, the exclusions relate to episodes that are irregular and may not be well-explained by the regression.

66. Specification (13) shows that by including all treatments in the analysis and not just the eight focal treatments—increasing the sample size to 46,390—has the effect of decreasing the magnitude of the estimated price-concentration relationship. As with specification (12), the model fits worse than the baseline regression. The statistical significance of the estimate is also lower, at 5 per cent. The worse model fit and lower level of statistical significance indicates that the regression explains the variation in the price data less well when there is a more heterogeneous mix of treatments included.

67. The final specification in Table 9, specification (14), shows that using more granular regional dummies, at a NUTS3 level rather than NUTS2, does not substantially change the results of the baseline regression. The estimated price-concentration relationship increases in magnitude to –0.204.[36]

68. Looking now at Table 10, estimation results are reported for the baseline specification but using data from each hospital group separately.[37] The results indicate some differences in the estimated price-concentration relationship when each set of data is used separately. The estimates range from 0.091 to –0.470. Of the five separate datasets, only when using the Nuffield dataset is the estimated relationship statistically significant (–0.470 at the 0.1 per cent level of statistical significance). Looking in more detail at results based on the other four datasets, it can be seen that the price-concentration relationships are estimated with low precision—each estimate is insignificant at the 5 per cent level, and the standard errors (not shown in the table) are 0.160 (BMI), 1.180 (HCA), 0.128 (Ramsay) and 0.078 (Spire). The 95 per cent confidence intervals associated with the majority of

[36] This specification to an extent also addresses the points discussed earlier regarding endogeneity. Using more granular regional categorization may mitigate any within-region differences in market condition and the potential effects of any endogeneity bias. The results of this specification are consistent with the findings from the endogeneity assessment—any bias may be relatively small and increase the magnitude of the estimated relationship.
[37] The eight focal treatments remain pooled together.

these imprecise estimates include the estimated price-concentration relationship from the baseline specification.[38]

69. One factor that will, in part, drive the different results when using the separate hospital group datasets is the ability of the regression methodology to identify a price-concentration relationship. This will differ according to the features of the dataset used. Reasons for this include differences in quantity or quality of data for each hospital group (eg HCA and Ramsay have fewer episodes available in our hospital dataset) and, perhaps most importantly, the differences in the portfolio of hospital sites contained in each dataset. The latter is relevant because, as noted early on in this paper, the regression analysis effectively makes comparisons between the prices charged at different hospital sites. By looking in isolation at only those hospital sites in the single hospital group datasets, there are necessarily less comparisons available than in the pooled dataset that contains data from all hospital groups (and their hospital sites).[39] More precisely, it is the variation in the LOCI measure between hospitals that identifies the price-concentration relationship, and this variation will be reduced in the single hospital group datasets as compared to a pooled dataset that includes all hospital groups.[40] It is therefore expected that there would be differences in the results when using only single hospital group datasets, and also expected that the estimates with these datasets would be less precise.

70. Pooling together the data of different hospital groups avoids these issues, by drawing on all of the data together which allows for comparisons across all hospital sites (and their associated LOCI measures). As a result we consider estimated relationships from the pooled dataset to be more reliable than those estimated using

---

[38] The baseline estimate of –0.180 lies marginally outside the lower 95% confidence interval (–0.176) when using the Ramsay dataset..
[39] In addition, the available comparisons in each dataset will differ.
[40] In an extreme situation, if all hospitals had the same LOCI there would effectively be no comparisons available to inform how prices may vary with LOCI, and the price-concentration relationship would not be identified.

the single hospital group datasets. Moreover, estimates using the pooled dataset are more informative of the price-concentration relationship of interest, that which operates at a broad level across treatments and the market as a whole. We are therefore of the view that it is appropriate to focus on the pooled dataset.

Conclusions

71. ToH1 argued that hospital groups may have market power in certain local areas, and that this may lead to adverse outcomes for consumers. The AIS set out our analysis showing that several hospitals appear to have local market power. The PCA, first discussed in the AIS and in more detail here, has been conducted to test whether local market power leads to higher prices for self-pay patients. This paper has described the PCA methodology, results and an analysis of the robustness of those results. The analysis indicates that there is a price-concentration relationship and that self-pay patients typically pay higher prices in more concentrated local areas.

72. The baseline specification, that consider eight focal treatments and uses LOCI as the concentration measure, indicates that a change in LOCI from zero (monopoly) to one (perfect competition) would result in a 18 per cent price reduction on average. More moderate changes in the LOCI are estimated to lead to more moderate price reductions. For example, an increase in the LOCI of 0.2 (corresponding to a decrease in average market share of 20 per cent) is estimated to lead to an average price reduction of around 3.6 per cent. The estimated relationship is statistically significant at the 0.1 per cent level.

73. A range of alternative specifications have also been considered. In general these alternatives provide support for the estimated relationship using the baseline specification—increases in LOCI are estimated to cause price reductions in most cases, and these estimated price reductions are comparable in magnitude.

Specifications using the fascia count variables indicate a price-concentration relationship that is weaker (ie of lower magnitude) and statistically insignificant; however, this is a likely consequence of fascia count being a less refined measure of local market concentration as compared with LOCI. We therefore intend to place more weight on the LOCI specifications. Other specifications that use LOCI as the concentration measure indicate statistically significant price-concentration relationships that imply price reductions typically in the 10 to 30 per cent range for a one-unit change in the LOCI. Estimating relationships at a more granular level, for instance on a treatment-by-treatment level, does highlight some differences amongst these more micro-relationships, but the differences are not of a magnitude that is thought to undermine the general findings described above.

74. In summary, the price-concentration relationship is thought to be relatively well characterized by the baseline specification. This indicates that reductions in local market concentration, as measured by LOCI, would likely lead to price reductions. The size of any price reduction depends on the change in LOCI. This estimated relationship appears robust to various alternative specifications.

Data processing

1. This appendix provides details of the data cleaning that has been undertaken to construct our two datasets for analysis—the hospital dataset and the Healthcode dataset.[41]

2. In both cases, information has been provided to us in the form of row-by-row invoice data. This means that each row in the data corresponds to a patient's purchase of a single item or service from a hospital. During a single hospital visit (an 'episode') a patient may receive many such items or services and therefore the data contains many rows of information for each episode. Across the different datasets we have received there are no standardized descriptions or codes available for each hospital item or service provided, and in some datasets, only the total price for all items and services received was available (ie the line item prices are not available). Our data cleaning process has therefore sought to standardize the definitions of the variables across each dataset, and consolidate the information to a level of aggregation where each row corresponds to a definition that is consistent across datasets.

3. We have consolidated the data to an episode-level, where an episode is defined as a single patient visit. In the data this is defined as a unique combination of patient identifier-discharge date-visit type-package indicator-date of birth-gender. The final datasets contain one row per episode, with aggregated information relating to that episode (eg the type of visit, the treating hospital, the particular treatment that was received, the primary specialty of the treating consultant, and the total episode price paid for all hospitals services). In principle each episode should correspond to a particular treatment and the primary specialty of the treating consultant. These two

[41] There have been minor changes to the hospital dataset since the AIS, which accounts for the increased number of observations that are referred to in this appendix relative to the AIS.

dimensions—treatment and specialty—are how we classify the data for most of our analyses. The key variable that has been created in this process is the episode price. This is the total price paid by a patient for all hospital services received during that episode. It excludes consultant fees and ancillary services; to remove these items we have followed advice given to us by the parties.[42]

4. During the process of consolidating the data we have noticed certain irregularities in the data. For example, episodes had missing information, episodes with admission dates occurring after discharge dates, and prices that were either unrealistically low or unexpectedly high. We have therefore applied a number of filters to the datasets in order to remove these irregularities so that they do not in any way distort our analysis. We first describe the filters applied to the both datasets (the hospital dataset used for the price variable, and the Healthcode dataset used to construct catchment areas and the LOCI). We have made exclusions for the following reasons:

(a) package episodes for which we could not identify the relevant consultant fee to remove (referred to below as 'package without part 2');

(b) package episodes for which there were inconsistencies in the price information between the two data sources submitted by hospital groups ('part 1 and part 2 inconsistencies');[43]

(c) episodes with admission dates occurring after discharge dates ('date inconsistencies');

(d) episodes with missing information for any of the following variables: patient identifier, type of visit, discharge date, package indicator, hospital postcode, gender, age ('missing data'); and

---

[42] In the case of consultant fees for non-package deals, the consultant fees were simply removed from the data before summing the cost of hospital services; for package deals, the consultant fees were extracted from the total package price using 'Part 2' of the DQ. In the case of ancillary services, where possible, these were removed from the row-by-row invoice data before summing the costs of other hospital services.

[43] Hospital groups submitted 'part 1' data and 'part 2' data. The former contained the prices for hospital services, and the latter contained invoices relating to consultant fees. For certain episodes both part 1 and part 2 contained prices for hospital services, and we have excluded episodes where the price of hospital services reported in part 1 and part 2 did not match.

(e) episodes with negative or zero episode prices.

5. After making these exclusions, we have then limited the data to the episodes that our analysis focuses on. This means excluding outpatient or day-case episodes, episodes relating to specialties outside of the 16 specialties and oncology, episodes for non-clinical treatments, episodes outside of the period 2009 to 2012, and episodes at hospitals outside of the 223 selected hospitals. These exclusions are collectively referred to as 'irrelevant data'.

6. Table 11 below shows the number of exclusions made to the data for each category.

TABLE 11 **Cleaning of the hospital datasets**

	BMI	*HCA*	*Nuffield*	*Ramsay*	*Spire*	*Healthcode*
Total episodes	1,404,122	550,238	933,968	59,062	940,902	14,566,178
Package without part 2	83,973	0	184,424	8,813	52,587	0
Part 1 and part 2 inconsistencies	322	0	0	56	0	0
Date inconsistencies	55	0	0	0	19	78,816
Missing data	10,368	652	7,199	22	5,652	2,062
Negative or zero prices	76,767	165,785	18,013	2,358	118,021	41,626
Irrelevant data	1,193,558	376,508	697,476	38,538	728,022	13,851,784
Total episodes after cleaning	39,079	7,293	26,856	9,275	36,601	591,890

Source: CC analysis.

Note: Numbers may not sum due to rounding. Exclusions are sequential, from the top to the bottom of the table. There were also a small number of exclusions made to the data following early discussions with parties; these exclusions are not shown in Table 11 (ie the 'Total episodes' figure is after these initial exclusions).

7. The cleaned hospital and Healthcode datasets therefore have sample sizes of 119,030 episodes (the sum of episodes from five operators' data) and 591,890 episodes over the period 2009 to 2012, respectively. The former relates to episodes for self-pay patients and the latter for insured patients. These are the samples of data used to create the catchment areas an LOCI measure.

8. The final stage of data preparation relates only to the hospital dataset and the PCA. In examining the price data for such episodes, we noted wide variation in the prices charged, even when evaluating episode prices for a single treatment at a single

hospital site. Some of this price variation is expected (eg due to differences in prosthesis or differences in patient requirements during a long hospital stay) but at least some of the variation is driven by factors that may potentially distort our analysis. Examples of factors that could cause this type of variation include IT, accounting or recording practices (eg refunds, data entry errors, cross-invoice recording) and particularly unusual patient circumstances (eg very complex episodes requiring multiple treatments). We have also sought to remove episodes that we cannot categorize to one particular treatment (ie CCSD code). We have therefore made the following exclusions:

(a) episodes with missing CCSD codes (referred to below as 'missing CCSD');

(b) episodes with invalid or more than one CCSD code ('invalid CCSD');

(c) irregular episodes, defined as either: episodes with a CCSD codes performed by a consultant with an atypical primary specialty; [44] episodes with a CCSD code that is uncommon in the data for a particular operator;[45] episodes with a low price that is less likely to be credible;[46] or, episodes with prices that appear extreme.[47]

9. Table 12 below shows the number of exclusions made to the data for each category.

[44] For the majority of treatments, a single primary specialty is common in the data (eg if the treatment is hip replacement, the specialty is typically 'Trauma and Orthopaedics'), but some instances an alternative primary specialty is listed. We have excluded episodes with these less common primary specialties.

[45] Episodes associated with operator-treatment combinations that have less than 30 observations in the data. (In the AIS we had previously applied this rule to hospital site-treatment combinations.) The main purpose of these exclusions is to ensure that the methodology for making exclusions relating to low or extreme prices can be applied more reliably. Both cases rely on making exclusions relative to the distribution of prices, and so if that distribution is based on a very small amount of data, it is difficult to determine with a systematic rule which parts of the data are 'extreme'. These episodes also represent a small minority of the data and are therefore not thought to be important.

[46] It is observed that certain episode prices observations lie very close to zero, or are very low relative to the majority of prices for that treatment. These episode prices observations likely contain some kind of discount, rebate or credit associated with them and are unlikely to represent the typical price for a particular treatment. We exclude such observations if they have an episode price that is less than 50 per cent of the median price for that treatment-operator combination.

[47] A price is considered extreme is if it less (or greater) than the lower (upper) quartile plus (minus) 1.5 times the inter-quartile range.

TABLE 12 **Cleaning of the hospital datasets**

	BMI	*HCA*	*Nuffield*	*Ramsay*	*Spire*
Total episodes after cleaning, excluding specialized hospitals	39,079	7,250	26,856	9,244	36,601
Missing or invalid CCSDs	11,657	1,057	7,225	2,182	7,428
Multiple CCSDs	7,642	0	0	2,296	6,383
Irregular episodes	6,758	4,805	5,359	2,118	7,439
Total episodes available for the PCA	13,022	1,388	14,272	2,648	15,351

Source: CC analysis.

Note: Numbers may not sum due to rounding.

10. The number of episodes available for the PCA is therefore 46,681 (the sum of the number of episodes for each operator).