CMA
Competition & Markets Authority

Estimating diversion ratios of hospital mergers

Economics Working Paper

September 2018

Estimating diversion ratios in hospital mergers

Cecilia Rossi[1]

Russell Whitehouse

Alex Moore

[1] Corresponding author (cecilia.rossi@cma.gsi.gov.uk). All authors are at the Competition and Markets Authority (CMA), London, UK. All views are those of the authors and not the CMA. We are grateful for helpful advice and comments throughout the project from Kate Collyer, Pasquale Schiraldi and Walter Beckert. We also thank Julie Bon, Chris Jenkins, Tom Kitchen, Paul Reeve, Laura Rovegno, Mike Walker and participants at the CMA Lunchtime Economics Seminars and the Competition in Hospital Markets conference in Rotterdam for helpful comments.

Abstract

Understanding patient choice is vital in assessing the closeness of competition between hospitals. The standard technique used in the UK is to estimate substitution patterns based on historical GP referrals. In this paper we compare the results of the 'GP referral' methodology to a demand estimation approach. Using patient-level data over a 3-year period (2012/13 – 2014/15) we apply both methodologies to every hypothetical merger between hospitals in England, for three specialties. We find a high degree of consistency between the two approaches, suggesting that GP referral analysis is a useful and reliable filter in merger cases. There are a small number of cases however in which the GP referral approach filters out potentially problematic mergers. Filtering should therefore be done with caution and in conjunction with additional evidence.

1. Introduction

Since 2006, National Health Service (NHS) patients have been free to choose which hospital to attend for their planned (elective) care, having first received a referral from their general practitioner (GP). Hospitals are paid a fixed fee for each patient, and so attracting more patients increases hospital revenues. As treatment is free for patients, hospitals – at least in theory – compete on quality to attract patients.

By reducing the extent of local competition, hospital mergers may therefore reduce the quality of patient care. It is the responsibility of the Competition and Markets Authority (CMA) to assess the likely extent of any reduction in competition which may give rise to a fall in quality, and balance this against possible countervailing factors.[2] Since 2012, the CMA and its predecessor organisations have undertaken a detailed ('phase 2') assessment of three NHS hospital mergers. Two of these were ultimately cleared, and one was prohibited.[3]

In assessing these mergers, a central question is how closely the merging hospitals compete for patients. Intuitively, if hospitals are close substitutes for one another, then their ability to reduce quality post-merger is greater: if a hospital reduces its quality then it risks losing patients, but a large share of these patients will be 'recaptured' by the other hospital. The *diversion ratio* captures the size of this effect – it quantifies the fraction of patients that would substitute to the other hospital if one of the merging hospitals were to lower its quality.

In this paper we compare two approaches to estimating the diversion ratio in hospital mergers. The first approach is to estimate a measure of closeness of competition based on historical market shares. This is done in the UK using an approach known as 'GP referral analysis', which estimates local market shares, and therefore diversion ratios, based on the number of patients referred by GPs to each hospital.

At heart, the GP referral methodology is intuitively appealing, and is a useful practical tool for merger analysis.[4] As hospital mergers are assessed on a specialty-by-specialty basis, with each specialty considered to be a separate economic market, GP referral analysis can be easily applied to each specialty to filter down

[2] If the CMA finds that a 'significant lessening of competition' (SLC) is likely to result from the proposed merger, it is required to consider the impact of any potential remedies on the 'relevant customer benefits' (RCBs) that might result from the merger. See Merger Remedies: Competition Commission Guidelines.

[3] The CMA has also considered four additional 'phase one' cases which entail a 40 working day review. Each case was cleared.

[4] The technique has historically been used only in hospital mergers. This appears to be because the data exists in a convenient form already for such cases, and because hospital choice differs from many other markets in that GPs and patients may together decide where the patient receives treatments and therefore a closeness of competition measure which accounts for GPs' role is intuitively appealing.

those that require detailed assessment.[5] On the other hand, inferring diversion ratios, and therefore closeness of competition, purely from historical market shares may be misleading, relying as it does on a number of implicit assumptions.[6]

An alternative approach is to estimate the diversion ratio using demand estimation. This econometric approach directly models patients' choice of hospital. In principle, this allows for more flexible substitution patterns. Demand estimation is generally more resource intensive however, and its use in a merger case would likely impose costs on the merging parties (for example in hiring external consultants to check the analysis and run sensitivities). The econometric approach also lacks the simple intuitive appeal of GP referral analysis. An important question is therefore whether GP referral analysis does a 'good enough' job at estimating diversion ratios to be used as a reliable filter in merger cases.

To test the referral analysis approach, we simulate the closure of each hospital in England, and calculate the diversion ratio from each hospital to both its closest competitor and to all hospitals within 50 km. This enables us to compare the two approaches on an aggregate basis. Our results show that the two approaches produce largely consistent (and highly correlated) estimates.

In particular, we consider whether each approach would 'filter in' or 'filter out' a merger for further consideration based on a threshold of a 40 percent diversion ratio. This threshold has been used in recent CMA cases to flag potentially problematic specialties for further assessment (e.g. CMA 2017). We find that the vast majority of cases would be consistently filtered in or filtered out under either methodology based on this threshold.

However, our results also show that a small number of cases would be filtered out under GP referral analysis, but filtered in under a demand estimation approach. This could mean that some potentially problematic cases would be cleared under a strict application of the GP referral analysis. This highlights that the GP referral analysis should be applied with some caution, and in conjunction with other evidence.

The paper is structured as follows. Section 2 presents the GP referral and demand estimation methodologies. Section 3 discussions the application, including both the data and the hypothetical merger case. Section 4 presents the results and section 5 concludes.

---

[5] Where there are limits to supply side substitution within specialties, markets may be defined more narrowly than a specialty. Outpatient and inpatient activities are also typically treated as separate markets (our data in this paper deals with inpatient activity). Markets are typically no wider than a specialty. See the CMA guidance on the review of NHS mergers for details.

[6] This includes the independence of irrelevant alternatives assumption, which is discussed further below.

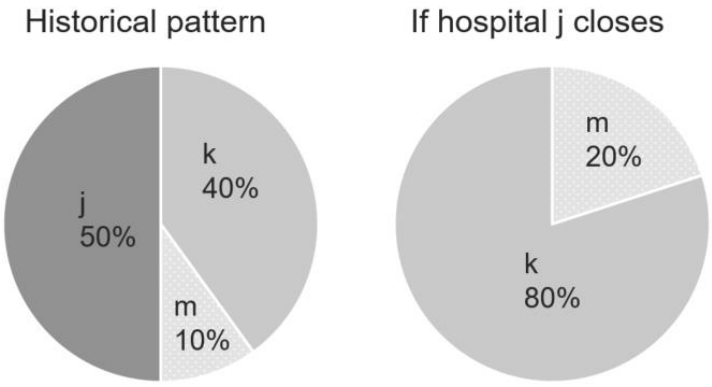
2. Methodologies

2.1 GP referral analysis

Broadly speaking, GP referral analysis estimates diversion ratios based on historical 'market shares'. Suppose for example that we are interested in the diversion ratio from hospital j to hospital k. The first step is to identify the set of GP practices that referred patients to hospital j (the 'anchor hospital') over a certain timeframe. This is typically set at 3 years.

For each GP practice, we then calculate the market share of each hospital based on patient referrals over the period. As in Figure 1, suppose that a particular GP practice referred patients to three hospitals – j, k and m. To calculate the diversion ratio from j to k, we 'close' hospital j, and re-calculate the implied market shares of k and m based on the ratio of their historical shares. For this GP practice, the j to k diversion ratio is therefore 80 percent. The overall j to k diversion ratio is calculated as a weighted average (based on patient referrals) across the set of GP practices identified above.

Figure 1: calculating diversion ratios using GP referral analysis



A major appeal of GP referral analysis is that it provides a relatively quick and intuitive indicator of closeness of competition between hospitals. In a practical merger case, this can be particularly useful in narrowing down a list of specialties for more detailed analysis. This is done by running the GP referral analysis on a specialty-by-specialty basis and filtering out those where the implied diversion ratios between the hospitals are below a chosen threshold; this has been set at 40 percent in some previous cases (e.g. CMA 2017). Further analysis is only given to those

which fail the filter, or those which are at the borderline, unless there are case specific reasons to undertake analysis of other specialties.[7]

2.2 Demand estimation

Our demand model closely follows that of Gaynor et. al. (2016). In particular, we specify a random utility model in which patient i chooses hospital j to maximise utility, specified as:

$$\begin{aligned} u_{ij} &= \bar{u}_{ij} + \varepsilon_{ij} \\ &= \gamma_i dist_{ij} + x_j'\beta + \xi_j + \varepsilon_{ij} \end{aligned} \tag{1}$$

where $dist_{ij}$ is the distance between patient i and hospital j, x_j is a vector of observable 'quality' variables for hospital j, ξ_j is a fixed effect for hospital j and ε_{ij} is an error term. Assuming that the error term is i.i.d. extreme-value distributed, equation (1) can be estimated as a conditional logit model.

We allow the impact of distance on hospital choice to vary across individuals. It may be the case for example that very sick patients are prepared to travel further. We do this by specifying γ_i as:

$$\gamma_i = \gamma + v_i \tag{2}$$

where γ is a fixed coefficient and v_i is a vector of observable characteristics for patient i. Allowing for patient heterogeneity gives the model much greater flexibility, which can produce more realistic elasticities and diversion ratios. In particular, the model no longer implies that substitution patterns are determined solely by market shares (we discuss this point in further detail in section 2.3 below).

Under the relevant statistical assumptions on ε_{ij} specified above, the conditional probability that individual i chooses hospital j is given by:

$$s_{ij} = \frac{\exp(\bar{u}_{ij})}{\sum_{m=0}^{M} \exp(\bar{u}_{im})} \tag{3}$$

[7] In merger cases, extensive integrity checks are conducted on the referral data, often in conjunction with clinical and coding expertise from the Parties. This is necessary because, although the underlying data is on average a good reflection of the treatments received by patients, coding practices vary in consistency across (and occasionally within) the same area. For example, for pragmatic reasons, the CMA might decide to analyse two specialties together because the same activity is coded under different codes by key parties. We do not undertake such checks in this paper, but rely on the data being coded 'on average' in a reasonably consistent way.

where M is the set of all hospitals. Summing across individuals, the *predicted* market share of hospital j can therefore be calculated as:

$$\hat{s}_j = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\exp(\hat{\bar{u}}_{ij})}{\sum_m \exp(\hat{\bar{u}}_{im})}\right) \quad (4)$$

where N is the total number of patients, and $\hat{\bar{u}}_{ij}$ is calculated using the coefficients from a logit regression based on equation (1) using patient-level data.

The diversion ratio between hospitals j and k – i.e. the fraction of j's patients diverting to hospital k in the event of a closure of j – can be calculated as:

$$D_{jk}^{c} = \frac{\hat{s}_{k,post} - \hat{s}_{k,pre}}{\hat{s}_{j,pre}} \quad (5)$$

where $\hat{s}_{z,post}$ is the estimated market share of hospital z in the period after the closure, and $\hat{s}_{z,pre}$ is the estimated market share of z in the period before the closure.[8]

As in GP referral analysis, equation (5) produces what is technically known as a 'closure diversion ratio'. That is, the formula shows the fraction of patients diverting to each hospital in the event that a particular hospital is closed. An alternative formulation of the diversion ratio is to assess the fraction of patients diverting to each hospital in response to a *marginal* change in price or quality (rather than the extreme case of closure). In such case, the diversion ratio will be calculated based on demand elasticities. We present the calculations for these elasticity diversion ratios in Appendix 1, and show that our results are not sensitive to using these (rather than closure diversion ratios) as the comparator.

2.3 Comparing implied diversion ratios

GP referral analysis assumes that, upon the closure of a hospital, patients substitute to the remaining hospitals in proportion to their historical market shares. By construction, it therefore imposes the 'independence of irrelevant alternatives' (IIA)

[8] The predicted market share post-merger ($\hat{s}_{k,post}$) can be calculated using the following formula:

$$\hat{s}_{k,post} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{s}_{ik}}{1 - \hat{s}_{ij}}\right)$$

where $\hat{s}_{ij}$ and $\hat{s}_{ik}$ are defined in equation (3).

property. This property implies that the relative probabilities of choosing any two hospitals are constant, regardless of the existence of alternative hospitals.

In some circumstances this property may be reasonable, but in other situations it is likely to fail. In the case of hospitals, IIA might not hold for example if some hospitals were more similar along a particular salient characteristic than others. We would then expect greater than proportional substitution between these hospitals, and lower than proportional substitution to others. In this case, market shares might not be a good guide to substitution patterns.

If the IIA assumption does not hold, then GP referral analysis will produce misleading results. Our demand model, by contrast, imposes much weaker assumptions, as we allow for individual patient heterogeneity. In particular, IIA need only hold within each 'group' of patients – i.e. those with the same characteristics – but does not hold across all patients in the same GP practice, or in the population.

Under the demand model approach, predicted patterns of substitution therefore vary across groups of patients, and the overall pattern of substitution is no longer determined simply by market shares. Allowing for patient heterogeneity in the model therefore allows for much more flexible substitution patterns – and potentially more accurate diversion ratios – than GP referral analysis. The trade-off is having a more complicated model.

The key question of interest here is whether the results of the referral analysis are 'good enough' to be used as a broad filter. We turn to that question in the following sections.

3. Application

3.1 Data

We have access to three years of patient-episode level data (2012/13 – 2014/5) from the Hospital Episode Statistics (HES).[9] HES is the data underlying the payments system for activity-based reimbursements to hospitals.[10] HES includes data on both elective and emergency treatments, although we focus our analysis here on elective treatments given that patient choice is not generally a consideration in non-elective care.

An episode is defined as a single period of care for a patient under one consultant; a patient's stay in hospital (a 'spell') may include several episodes. For each spell in the dataset, we select the most relevant episode, which typically corresponds to the most resource intensive episode.[11] This is to ensure that we are studying the event that was most likely to be relevant when the patient made their choice of hospital. Other episodes within a given spell may be caused by complications that occurred during the main episode for example.

For each episode, the dataset identifies the treatment hospital and site and the GP practice that referred the patient. Following standard practice, we define a hospital as an NHS 'trust', and each hospital can consist of several sites (typically close to one another). When calculating distance, we use the patient's distance to the actual site at which they were treated. On average, there are around 2 sites per hospital (see table 1).

The dataset includes rich information on the patient's treatment whilst in hospital, and their individual characteristics. This includes the medical specialty (e.g. urology) and a number of patient characteristics such as age, gender and the number of

[9] We are grateful to NHS Digital for providing us with an extract of HES data under contract reference DARS-NIC-38368-V3S5C-v2.12 for the purposes of undertaking research into merger control in the NHS, specifically including the extent to which choice is exercised at various nodes along the patient pathway and the implications of this for merger assessment. The data is subject to copyright: Copyright © (2018), the Health and Social Care Information Centre. Re-used with the permission of the Health and Social Care Information Centre. All rights reserved.

[10] HES is an extremely large and comprehensive dataset. An extract of HES was provided to the CMA for the purposes of undertaking research into merger control methodologies. We undertook a 'cleaning' exercise based on several steps. In doing so, we exclude: patients whose date of birth is unknown; patients who had more than 5 episodes within the spell; patients whose location is unknown or live outside England; patients who attend private hospitals; patients who attended 'unknown' or 'unverified' sites; regular attenders and day cases.

[11] We select the relevant episode based on which episode of care the hospital would be paid for. We do this by using an established 'spelling methodology' which, by reviewing diagnosis and procedure codes of each episode of care, allows us to identify which is the main episode in each spell. This typically coincides with the most resource intensive.

comorbidities. The number of comorbidities enables us to classify the 'severity' of the patient's condition, which we enter into the demand model as a binary variable. We also have data on the 'lower super output area' that each patient lives in, which allows us to (i) calculate the distance from the patient's residence to the hospital, and (ii) proxy for the socio-economic characteristics of the patient.[12]

Throughout the analysis, we focus on three specialties: urology; trauma and orthopaedics; and ear, nose and throat (ENT). As noted in the introduction, in a merger case each specialty is typically considered to be a separate market, and the GP referral analysis is applied as a filter on a specialty-by-specialty basis. We therefore follow this approach here and consider each specialty individually. Our three specialties were chosen as they are relatively large, predominantly elective and generally well-coded in the dataset. They represent the type of care that is typically given in different hospital departments, and are specialties in which the CMA has previously considered whether mergers could give rise to SLCs.

For the demand model, we combine HES data with hospital 'quality' indicators from NHS Digital. Due to the hospital fixed effects in the model, we focus only on those indicators which vary within a hospital over time and are plausibly exogenous. The first quality indicator is the Summary Hospital Level Mortality Indicator (SHMI or 'mortality rate'), which is available annually on a hospital-level basis. The variable is calculated as the ratio of the observed number of deaths to the 'expected' number of deaths at each hospital, and therefore captures 'excess mortality'; we expect an increase in this variable to reduce the probability that a patient chooses a particular hospital.[13] The second variable is the number of medical staff at each hospital, which is available monthly on a hospital-level basis.[14] We expect an increase in this variable to increase the probability that a patient chooses a particular hospital.

Summary statistics are presented in table 1. For each specialty there are around 180 hospitals, and between 300 and 400 sites. In each case patients typically travel around 10 km to their chosen hospital. There is considerable variation across the three specialties however in terms of patient profiles. In particular, patients are typically much older, and with more severe conditions, in urology than ENT; ENT

[12] Distance is calculated as the straight-line distance between the centroid of the super-output area (coordinates available from the Office for National Statistics) and the location of the hospital site that the patient attended (coordinates available from the Office for National Statistics, based on the postcode of the hospital). Data on the socio-economic characteristics of each super-output area are available via the Office for National Statistics.
[13] The average SHMI across all hospitals is 1.0. By accounting for the 'expected' number of deaths, the variable controls for the fact that different hospitals serve different types of patients. The numerator is the total number of finished provider spells for each trust which resulted in a death either in hospital or within 30 days (inclusive) from the patient discharge. The denominator is computed using a risk-adjusted model with a patient case mix of age, gender, admission method, year index, Carlson Comorbidity Index and diagnosis grouping. The model is estimated on a three-year dataset.
[14] We use full-time equivalent medical staff. The average number of medical staff at each hospital is 347.

patients are also more likely to come from a low-income area. This variation is helpful in ensuring that we are testing the two empirical methodologies across a diverse set of specialties.

Table 1: Summary statistics

	Urology	Trauma & orthopaedics	ENT
Hospitals (N)	180	178	188
Sites (N)	393	397	314
Patients (N)	400,338	733,220	223,235
Distance (km)	9.4	9.6	9.7
Age (years)	66.6	60.2	36.8
Severity (%)	43.5	32.6	26.1
Low income (%)	49.2	48.9	56.9
Rural (%)	19.1	20.0	17.1

Unless specified, the table shows the mean for each variable across the three years of data (2012/13 – 2014/15).

4. Results

4.1 Demand estimation results

Before comparing the estimated diversion ratios of our two methodologies, in table 2 we first present the econometric results of the demand estimation. These coefficients will subsequently be used to compute the diversion ratios from the demand model (based on equation 5).

In each column, we find that the probability that a patient chooses a particular hospital declines rapidly and significantly with distance. This effect is greater (i.e. the coefficient is even more negative) for older patients, those on low incomes (except ENT) and those in rural areas. For urology and ENT, we find that the negative impact of distance on hospital choice is less pronounced for more 'severe' patients (i.e. those with a greater number of comorbidities).

The quality variables generally enter as expected. An increase in the mortality rate reduces the number of patients for two of the specialties, although the coefficient is positive (and significant at the 10 percent level) for ENT. It is not clear why the coefficient would be positive in this case, although we note that this is significant only at the 10 percent level. It is also the case that a number of the coefficients for ENT have the opposite sign to the other specialties, and so the results in general may be less robust for this specialty. In all three specialties an increase in the number of medical staff increases the number of patients, and the coefficient is significant in two out of three cases.

We emphasise however that caution should be used when interpreting the coefficients on the two quality indicators. Due to the use of hospital fixed effects, the coefficients are based purely on variation *over time* at each hospital. To the extent that such time variation is limited, the model may not pick up the true impact of the quality variables on hospital choice.[15]

[15] We further note that patients' may have highly varying pre-treatment health, and pre-treatment health could impact choice differently across specialties. Being older is likely to be associated with greater personal costs of travel, and therefore exacerbate the effect of distance, but it is also likely to be associated with underlying health complexities which give patients greater preparedness to travel for the right care. The 'net' effect could vary across specialties leading to interaction terms with different signs. Furthermore, if there is simultaneity between pre-treatment health and hospital mortality rates (such that sicker patients attend better hospitals, but better hospitals attract sicker patients), this could also cause the mortality rate to have an upwards biased coefficient. Although these particular coefficients may not be fully reliable, this does not affect our results in any meaningful way. In particular, the exclusion of these variables has little impact on the other coefficients in the model (distance and its interactions, and the hospital fixed effects) and therefore has little impact on the implied diversion ratios. We note that the general (time-invariant) impact of quality at each hospital is captured through the hospital fixed effects.

Table 2: Demand estimation results

	(1) Urology	(2) Trauma & orthopaedics	(3) ENT
Ln(distance)	-1.183***	-1.556***	-2.188***
	(0.019)	(0.011)	(0.017)
Standardised mortality rate	-0.383***	-0.110*	0.221*
	(0.092)	(0.064)	(0.134)
Ln(staff)	0.146***	0.178***	0.011
	(0.054)	(0.036)	(0.073)
Distance interactions			
Age	-0.018***	-0.006***	0.001***
	(0.00)	(0.00)	(0.00)
Low income	-0.057***	-0.058***	0.028*
	(0.01)	(0.006)	(0.014)
Severity	0.299***	0.002	0.126***
	(0.01)	(0.006)	(0.015)
Rural	-0.828***	-0.877***	-1.066***
	(0.016)	(0.01)	(0.025)
Site fixed effects	Yes	Yes	Yes
Observations	2,370,191	4,778,637	1,002,771
Pseudo-R2	0.52	0.44	0.49

Standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

As the logit model is non-linear, the coefficients presented in table 2 have no direct economic interpretation. To demonstrate the magnitude of the variables on patient choice, we have therefore computed elasticities for the coefficients on distance, standardised mortality rate and staffing levels. These are presented in table 3; full analytical derivations are included in Appendix 1.

For urology, we find that a 1 percent increase in the distance of a hospital would on average result in a 0.5 percent decrease in the probability that a patient selected a given hospital. That is, doubling the distance between the patient and hospital roughly halves the probability that the patient will attend that hospital: distance is therefore a key driver of patient choice. A 1 percent increase in the standardised mortality rate would lead to a 0.16 percent decrease in choice probability, and a 1 percent increase in the number of medical staff would lead to a 0.06 percent

increase in choice probability.[16] The elasticities for the main effects in the other specialties are interpreted likewise.

Table 3: implied elasticities

	(1) Urology	(2) Trauma & orthopaedics	(3) ENT
Distance	-0.50%	-0.77%	-0.94%
Standardised mortality rate	-0.16%	-0.05%	+0.09%
Staff	+0.06%	+0.09%	+0%

4.2 Implied diversion ratios: aggregate results

We now compare the implied diversion ratios of the demand model (based on the results in table 2) and the GP referral analysis. We focus on a particular specialty (urology), although the results for the remaining two specialties are very similar. We show results for all three specialties further below (figure 4).

We first compare the estimated diversion ratio between every hospital in England and its 'closest competitor' based on the two methodologies (figure 2). The closest competitor is the hospital for which the diversion ratio is the highest; this may differ depending on the methodology. As well as showing the overall correlation, figure 2 therefore also indicates whether the two methodologies identify the same closest competitor. A dark blue dot indicates that the same hospital is identified as the closest competitor; a light blue dot indicates that the closest competitor is 'ranked' only one place differently (i.e. first under GP referral analysis and second under the demand model or vice-versa); a yellow dot indicates that the closest competitor is ranked more than one place differently under the two methodologies.

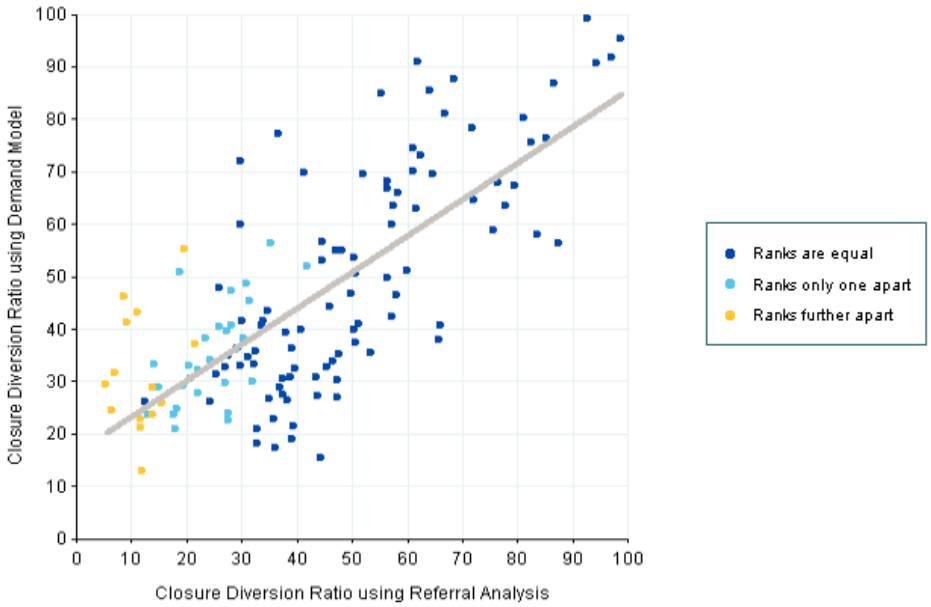
We can see in figure 2 that the two approaches produce largely consistent and highly correlated diversion ratios between each hospital and its closest competitor. There is a reasonable degree of variation however, and the overall R-squared is just 56 percent. This is particularly true for 'intermediate' cases in which the diversion

[16] There is some variation in the elasticities across hospitals. We calculated hospital-specific elasticities. For Urology we again found that in response to a 1% increase in the standardised mortality rate, some hospitals would lose over 0.35% of their demand, whilst others would lose less than 0.05%.

ratio is between 30 and 70 percent; in this range there is considerable variation around the line of best fit.

Although there is some variation in the diversion ratios produced by the two methodologies, the right side of figure 2 shows that the ranking of the closest competitor is largely consistent. We see that in all cases where the GP referral analysis finds a diversion ratio above 20 percent for example, the demand model identifies the same closest competitor or the ranking is only one place apart. Below the 20 percent threshold the rankings of the two methodologies are much less consistent (indicated by the yellow dots). Mergers with such low diversion ratios may be unlikely to raise significant competition concerns, although we note that in some of these cases in figure 2 the demand model finds a diversion ratio well above 40 percent.

Figure 2: implied diversion ratios, GP referral analysis and demand model



Of course, mergers do not necessarily occur only between the closest competitors. Mergers between any hospitals which are reasonably close together are also plausible and may result in competition concerns.[17] In figure 3 we therefore consider the set of all hypothetical mergers between all hospitals located within 50 km of each other.

For each potential merger, we indicate whether the estimated diversion ratio produced by the two methodologies is above or below 40 percent; a 40 percent

[17] Mergers between hospitals located a significant distance apart to create 'hospital chains' have not yet been a common feature of consolidation in the English NHS.

threshold has been used by the CMA as a filter in previous NHS merger cases (e.g. CMA 2017). The blue dots indicate those cases that would either have been 'filtered in' or 'filtered out' for further analysis under either of the two methodologies (based on the 40 percent threshold). It is clear from figure 3 that the majority of cases fall into this category, meaning that the two approaches generally produce consistent results.

The amber dots in figure 3 indicate those cases that would have been filtered in for further analysis using the GP referral analysis, but filtered out under demand estimation. These are potentially 'false positives' as these cases would be subject to detailed analysis under GP referral analysis, but in reality may not be a cause for concern. More concerning are the red dots – the 'false negatives' – as these cases would be filtered out by the GP referral analysis, but filtered in by the demand estimation. These cases may therefore be a cause for concern, but would not form part of a detailed merger assessment based on a strict application of the 40 percent threshold (and a strict interpretation of the GP referral results).

It is clear from figure 3 that a strict 40 percent filter results in a number of 'false negatives'. Most notably, in a small number of cases the demand model implies a diversion ratio well above 50 percent, and yet the case is still filtered out under GP referral analysis. This therefore suggests that a mechanical application of a 40 percent filter could result in some potentially problematic specialties being filtered out of the detailed merger analysis – i.e. incorrectly cleared.

However, the percentage of such case is relatively low. In figure 3, just 1.6 percent of cases are shown to be 'false negatives', and 2.1 percent are 'false positives'. Over 95 percent of cases are therefore filtered consistently using either GP referral analysis or demand estimation.[18] As it is the false negatives that we are particularly concerned about, these numbers are encouraging: our results suggest that GP referral analysis filters in the vast majority of cases that are likely to cause competition concerns for further analysis.

[18] The r-squared for this chart is 79%. In Appendix 1 we show that using elasticity- rather than closure diversion ratios improves the fit between the methodologies; these results are therefore on the conservative side.

Figure 3: Comparison of implied diversion ratios for hospitals within 50 km

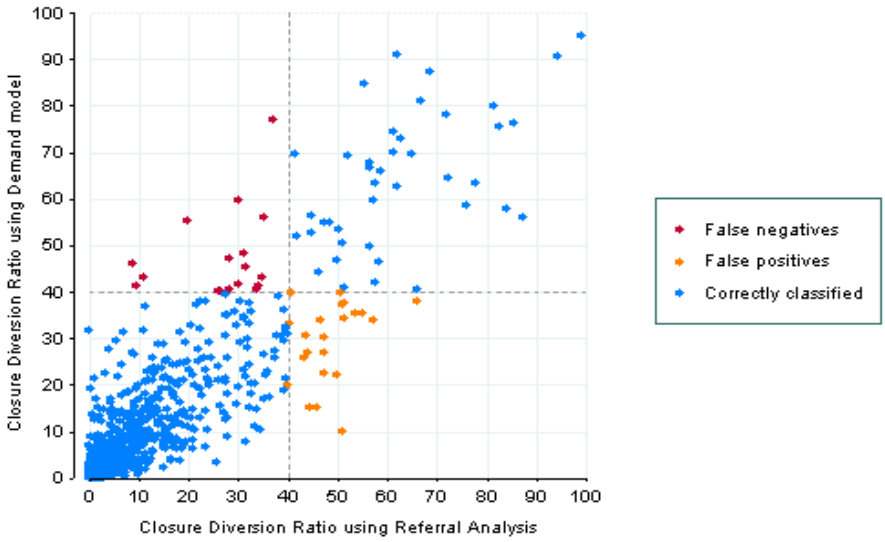
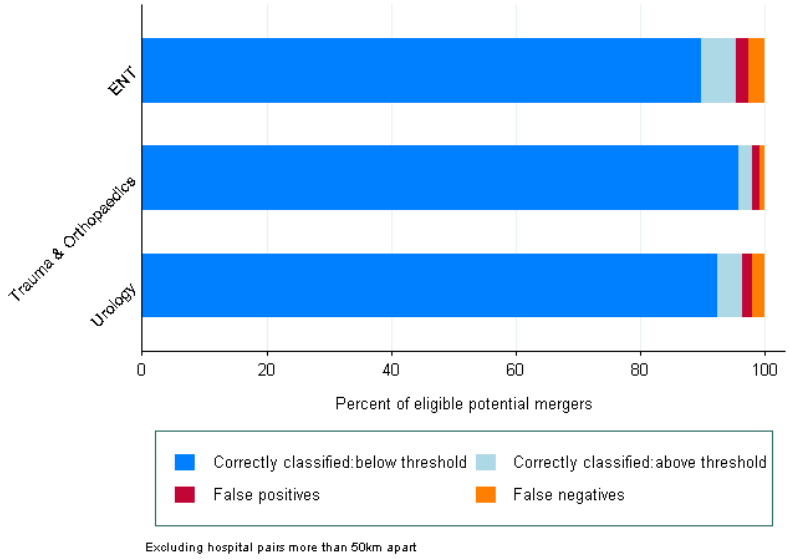


Figure 4 shows that similar results arise when we consider the two other specialties (ENT and trauma and orthopaedics). In particular, figure 4 shows that over 90 percent of hypothetical mergers between trusts within 50 kilometres are classified the same way under both the GP referral and demand methodologies, and that the ratio of false positive to false negatives is roughly equal in each case.

Figure 4: Classification consistency across methodologies for each specialty



From the perspective of filtering, it is most important to pick up all specialties which could cause competition concerns, and then to reduce this number down using other sources of evidence. As a consequence, false negatives are of greatest concern when filtering. Whilst the proportion of these results appears to be low, it would be

useful for a competition authority to be able to check whether specialties not flagged by the 40 percent filter in specific assessments are actually concerning.

To this end, we examined the characteristics of false negatives flagged in figure 3, as compared with ‘true negatives’ where both demand and referral analysis predict diversion of less than 40 percent. We found that 75 percent of the false negatives were between the geographically most-proximate alternatives. The equivalent figure for ‘true negatives’ was much lower. We also found that a higher proportion of false negatives were in rural areas than conurbations, and that they had lower rates of harm.[19]

The greatest concern would arise where the two methodologies give widely different results. This is not common: our analysis showed that in only two hypothetical mergers, the demand analysis predicted diversion of over 60 percent, but the GP referral analysis predicted diversion of less than 40 percent. In both cases, both methodologies identify the same trust as the closest competitor; the referral analysis however distributes more patients to the second and third alternatives. There may be good reasons for this: hospital referral patterns are complex and depend on specialisation at the local level which would not be captured in our relatively simple demand model. Further, our demand model uses straight line distance, which may not be a good proxy for travel time at the individual local level.

This supplementary analysis therefore supports the strength of referral analysis: provided that a competition authority does not implement a filter without regard for other evidence, it appears to classify hypothetical mergers correctly in almost all cases. Further, our methodology of assessing referral analysis’s strength may be conservative: some false negatives turn out to be potential ‘true’ negatives: as we show in Appendix 1, using elasticity diversion ratios rather than closure diversion ratios improves the consistency between the models.

[19] In one case, the demand analysis suggests a hospital located across a significant body of water is the closest competitor of the other: the referral analysis is able to use local referral patterns to avoid this pitfall!

5. Conclusion

This paper compares two alternative approaches to estimating diversion ratios in hospital mergers. The first, GP referral analysis, is a relatively simple and intuitive approach based directly on market shares. This approach has been used in a number of recent merger cases at the CMA, and is primarily a means of filtering down the list of specialties that will be taken forward for in-depth analysis. The second approach, econometric demand estimation, is more resource intensive and complex but allows for more flexible patient substitution patterns than GP referral analysis; and therefore, potentially, more accurate diversion ratios.

Given the practical appeal of GP referral analysis, the key question is whether it is 'good enough' to be used as a filter in merger cases. The results presented here show that there is a very high degree of consistency in the results generated by GP referral analysis and the demand model. Based on a threshold of 40 percent diversion, the two approaches would consistently filter in, or filter out, around 95 percent of cases for further analysis. Less than 2 percent of cases would be filtered out by GP referral analysis, but filtered in under the demand model. Our results therefore show that, overall, GP referral analysis is a useful and reliable filter in merger cases.

References

Beckert, Walter; Christensen, Mette; and Collyer, Kate, (2012) 'Choice and Competition in NHS-Funded Acute Services in England', *The Economic Journal*, 122 (560): 400-417.

Competition and Markets Authority, (2017) *Central Manchester University Hospitals and University Hospital of South Manchester: A report on the anticipated merger between Central Manchester University Hospitals NHS Foundation Trust and University Hospital of South Manchester NHS Foundation Trust.*

Domencich, Tom; and McFadden, Daniel L. (1996), 'Urban Travel Demand: A Behavioral Analysis' *North-Holland Publishing Co*: 83-87.

Gaynor, Martin; Propper, Carol and Seiler, Stephan, (2016) 'Free to choose? Reform, choice and consideration sets in the English National Health Service', American Economic Review, 106(11): 3521–3557.

Katz, M. and Shapiro, C. (2003), 'Critical Loss: Let's Tell the Whole Story', Antitrust, 2003.

Train, Kenneth *Discrete Choice Methods with Simulation*, (2009) Cambridge University Press.

Appendix 1

This appendix sets out additional formulae and explanations relevant to the statistics presented in the main paper.

Marginal Effects

In order to calculate the elasticities with respect to distance, mortality rate and staffing levels, we must first compute marginal effects. A marginal effect is the unit change in choice probability in response to a unit change in an explanatory variable.

These are first calculated at the individual level by differentiating predicted choice with respect to explanatory variables. We then calculate average marginal effects at the overall and individual level, as inputs for our elasticity formulae.[20] Following Domenich & Mcfadden (1996) and Train (1971), we derive these as follows:

$$M_{ij}^{v} = \frac{\partial \hat{s}_{ij}}{\partial V_j} \tag{A1}$$

$$= \frac{\partial \left((\exp(\hat{\bar{u}}_{ij}) \cdot (\sum_m \exp(\hat{\bar{u}}_{im}))^{-1} \right)}{\partial V_j} \tag{A1.2}$$

$$= \left(\frac{\partial(\exp(\hat{\bar{u}}_{ij}))}{\partial V_j} \cdot \left(\sum_m \exp(\hat{\bar{u}}_{im}) \right)^{-1} \right) + \left(\exp(\hat{\bar{u}}_{ij}) \cdot \frac{\partial (\sum_m \exp(\hat{\bar{u}}_{im}))^{-1}}{\partial V_j} \right)$$

by the product rule (A1.3)

$$= \left(\frac{\partial(\hat{\bar{u}}_{ij})}{\partial V_j} \cdot \frac{\exp(\hat{\bar{u}}_{ij})}{\sum_m \exp(\hat{\bar{u}}_{im})} \right) + \left(- \exp(\hat{\bar{u}}_{ij}) \cdot \frac{\partial(\hat{\bar{u}}_{ij})}{\partial V_j} \cdot \exp(\hat{\bar{u}}_{ij}) \cdot \left(\sum_m \exp(\hat{\bar{u}}_{im}) \right)^{-2} \right)$$

by the chain rule (A1.4)

$$= \frac{\partial(\hat{\bar{u}}_{ij})}{\partial V_j} \cdot \frac{\exp(\hat{\bar{u}}_{ij})}{\sum_m \exp(\hat{\bar{u}}_{im})} \cdot \left(1 - \frac{\exp(\hat{\bar{u}}_{ij})}{\sum_m \exp(\hat{\bar{u}}_{im})} \right)$$

[20] This is in contrast to calculating Marginal Effects at the Mean (MEM), where we would set all variables to their means before calculating margins. AME is to be preferred over MEM because it uses the entire distributions of the variables as they are in its calculation.

$$= \beta_v \cdot \hat{s}_{ij} \cdot (1 - \hat{s}_{ij})$$

where utility is linear in explanatory variable v (A1.6)

where subscript v denotes explanatory variable v, which may for example be distance or SHMI.[21]

Elasticities

We use these individual level marginal effects to calculate demand elasticities, i.e. the proportional change in own hospital demand following a proportional change in the values of a variable of interest.[22]

We calculate elasticities *for each hospital* by finding the sum of responsiveness of each individual's demand for that hospital to quality, weighted by the contribution of the individual to the overall demand for that hospital ($\frac{\hat{s}_{ij}}{\hat{s}_j}$), as follows.

$$E_j^v = \frac{\partial \hat{s}_j}{\partial V_j} \cdot \frac{V_j}{\hat{s}_j} \tag{A3}$$

$$E_j^v = \sum_i^I \left[\beta_v \cdot \hat{s}_{ij} \cdot (1 - \hat{s}_{ij}) \cdot \frac{\hat{s}_{ij}}{\hat{s}_{ij}}\right] \cdot \frac{V_j}{\hat{s}_j} \tag{A3.1}$$

$$E_j^v = \sum_i^I \left[\beta_v \cdot (1 - \hat{s}_{ij}) \cdot V_j \cdot \left(\frac{\hat{s}_{ij}}{\hat{s}_j}\right)\right] \tag{A3.2}$$

where E_j^v is the elasticity for hospital j with respect to explanatory variable v.[23] This elasticity is used to calculate elasticity diversion ratios at hospital level.

---

[21] In our case, utility is not linear in staff numbers or distance, but the log of these terms. In these cases, the partial derivative $\frac{\partial(\hat{u}_{ij})}{\partial V_j}$ is $\frac{\beta_j}{V_j}$, giving the formula for (individual level) marginal effects:

$$M_{ij}^v = \frac{\beta_v}{V_j} \cdot \hat{s}_{ij} \cdot (1 - \hat{s}_{ij}) \tag{A1.6.1}$$

[22] We use the elasticities for mortality rate, although it turns out that the elasticity does not depend on any specific beta terms or covariates.

[23] Where the explanatory variable is in logs, we could modify the expression above to give the following term

$$E_j^v = \sum_{i=1}^{N_j} \frac{\beta_v}{V_j} \cdot V_j (1 - \hat{s}_{ij}) \cdot \left[\frac{\hat{s}_{ij}}{\hat{s}_j}\right]$$

We also calculate the overall elasticity of demand as the sum of the individual level 'elasticities', weighted by each's contribution to overall demand.

$$E^v = \sum_i^I \beta_v \cdot V_j \cdot \left(1 - \hat{s}_{ij}\right) \cdot \left[\frac{\hat{s}_{ij}}{\hat{s}}\right] \tag{A3.3.1}$$

where E^v is the proportional change in overall demand with respect to explanatory variable v. These elasticities are reported in Table 3.

We have focussed so far on own-elasticities. However, cross-elasticities are also a component of elasticity diversion ratios. Cross-elasticities are the change in demand for one hospital k in response to a change quality at a different hospital j.

To find these, the first step is again the calculation of individual marginal effects; in this case, the unit change in the probability that each individual will select a given hospital k following a unit change in quality of j. The derivation is analogous to that given for own-elasticities above.

$$M_{ijk}^v = \frac{\partial \hat{s}_{ik}}{\partial V_j} \tag{A4}$$

$$= -\frac{\partial(\hat{\bar{u}}_{ik})}{\partial V_k} \cdot \frac{\exp(\hat{\bar{u}}_{ik})}{\sum_m \exp(\hat{\bar{u}}_{im})} \cdot \frac{\exp\left(\hat{\bar{u}}_{ij}\right)}{\sum_m \exp(\hat{\bar{u}}_{im})} \tag{A4.1}$$

$$= -\beta_v \cdot \hat{s}_{ij} \cdot \hat{s}_{ik}$$

where utility is linear in explanatory variable v (A4.2)

where M_{ijk}^v is the percentage point change in individual i's predicted choice probability for hospital j following a unit change in explanatory variable v at hospital k.

The second step is to turn the marginal effects into elasticities:

$$E_{jk}^v = \frac{\partial \hat{s}_k}{\partial V_j} \cdot \frac{V_j}{\hat{s}_k} \tag{A5}$$

$$E_{jk}^v = \sum_i^I \left[\beta_v \cdot \hat{s}_{ij} \cdot V_j \cdot \left(\frac{\hat{S}_{ik}}{\hat{S}_k}\right)\right] \tag{A5.1}$$

where E_{jk}^v is the elasticity for hospital j with respect to explanatory variable v.

Elasticity diversion ratio

Having calculated the hospital level own- and cross-elasticities, it is possible to compute diversion ratios. The diversion ratio from hospital j to hospital k measures the proportion of people who would switch from j to k, following a small change in quality at hospital j.

The formula for these diversion ratios is as follows:

$$D_{jk}^{P} = \frac{E_{jk}^{P} \cdot \hat{s}_k}{-E_{jj}^{P} \cdot \hat{s}_j} \qquad \text{(A6)}$$

In practice, several terms are common to the own- and cross-elasticity formulae: it turns out that almost all of these were introduced when we converted marginal effects into elasticities. We can therefore use just the ratio of the cross- and own-marginal effects summed across all individuals to compute the diversion ratios.[24]

$$D_{jk}^{P} = \frac{\sum_i^I [\beta_v \cdot \hat{s}_{ij} \cdot \hat{s}_{ik}]}{\sum_i^I [\beta_v \cdot \hat{s}_{ij} \cdot (1 - \hat{s}_{ij})]} = \frac{\sum_i^I [M_{ijk}^v]}{\sum_i^I [M_{ij}^v]} \qquad \text{(A6.1)}$$

$$D_{jk}^{P} = \frac{\sum_i^I [\hat{s}_{ij} \cdot \hat{s}_{ik}]}{\sum_i^I [\hat{s}_{ij} \cdot (1 - \hat{s}_{ij})]} \qquad \text{(A6.2)}$$

This formula is similar to that which is used to compute the 'closure' diversion ratios. In particular, like the formula for closure diversion ratios, it is simply a function of market shares.

Using elasticity rather than closure diversion ratios continues to show a high correlation with the results of the referral analysis. In fact, the correlation is somewhat higher,[25] and the false negative rate notably lower.

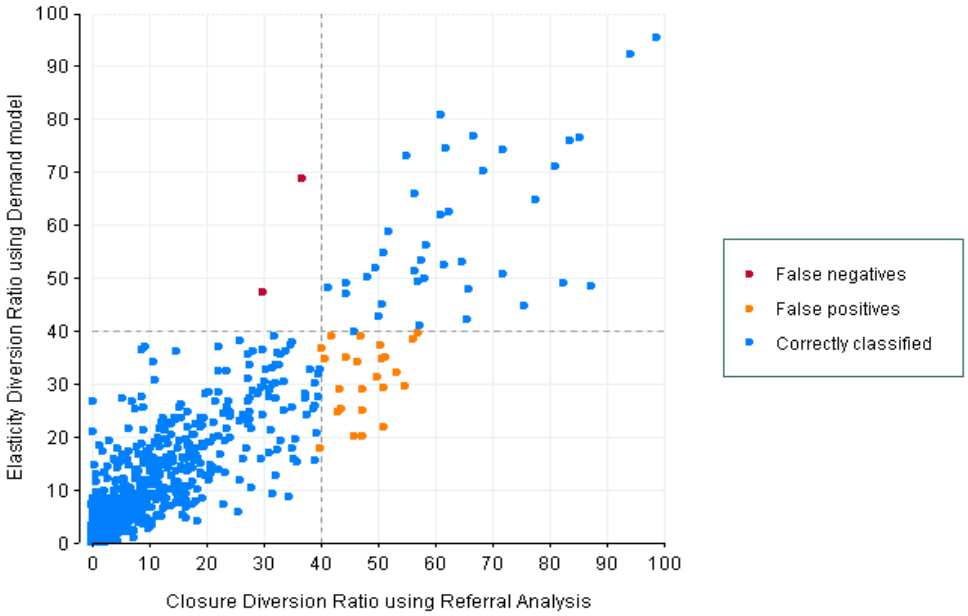
Figure A1 demonstrates this, using an amended version of Figure 3 above. This shows the relationship between the elasticity diversion ratio and the referral analysis diversion ratio for all hypothetical mergers in Urology within 50km.[26]

[24] We could even cancel the β_v, although we have retained them in the formulae to highlight the equivalence of the numerator to the cross-marginal effect and the denominator to the own-marginal effect.

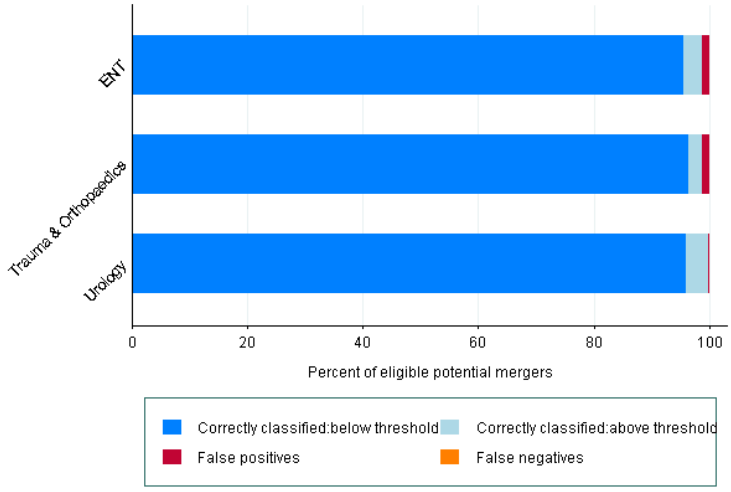
[25] The r-squared is 83%

[26] The picture is not dissimilar in other specialties, which we have omitted for brevity

Figure A1: Comparison of elasticity and referral analysis diversion ratios



Again, similar results hold true for other specialties, as shown in Figure A2 below



Predicting trust-level results

For simplicity in setting out the formulae above, we have assumed that all hospital trusts have only one site. In practice, many have at least two. To account for this, we simply aggregate market shares across sites.

$$\hat{s}_J = \hat{s}_{j1} + \hat{s}_{j2} \qquad (7)$$

Where $\hat{s}_J$ denotes the share of trust J, comprised of two sites $j1$ and $j2$.

In doing this, we are assuming that all sites experience a uniform change in quality arising from the merger. This could occur for elements of quality not set at site level (for example, quality parameters such as CEO quality might be equal across trusts).

We are also assuming that patients experiencing a quality drop at their preferred site would not switch to another site within the same trust. This might be plausible if patients experience a “brand reaction”. We note that websites such as NHS choices have historically presented the same quality measure for each site in multi-site trusts

Where quality variables are set and perceived to be differentiable at site level, the site level elasticities (and diversion ratios) may be more appropriate. For simplicity, we have conducted analysis at trust level.