THE UNIVERSITY of York

CENTRE FOR HEALTH ECONOMICS

Interpreting the NHS Cost Indices for Acute Trusts

Andrew Street

DISCUSSION PAPER 175

INTERPRETING THE NHS COST INDICES FOR ACUTE TRUSTS

Andrew Street

Centre for Health Economics, University of York

December 1999

CENTRE FOR HEALTH ECONOMICS DISCUSSION PAPER SERIES

The Discussion Paper (DP) Series was established in 1984 to make research available quickly to a broad readership drawn from the academic community, professional health care staff and policy makers, both within the UK and internationally. Material published in DP form ranges from work in progress, to edited versions of completed project reports, from papers intended ultimately for journal publication, to those which seek to stimulate immediate debate. Hence the variation in length and complexity of DPs.

In the past, DPs have been subject to a minimum of editorial revision, it being principally seen as the responsibility of individual authors to ensure that appropriate standards were met. Since June 1996, a more formal system of internal peer-review has been introduced to consolidate the high levels of quality that have been achieved in the past. That system involves the refereeing of DPs in draft form, with any recommended revisions being verified independently by a senior member of staff.

DPs are not a substitute for publication in peer-reviewed journals. It is to be hoped that journal editors will recognise the nature of the DP series, and accept that this does not compromise the status of those papers published by the Centre which may later be submitted for consideration in more formal publications. Comments and suggestions on DPs are encouraged and should be sent directly to the authors.

ACKNOWLEDGEMENTS

The author would like to thank Gwyn Bevan, Diane Dawson, Maria Goddard, Karin Lowson, Rowena van der Merwe, Peter Smith, David Speigelhalter and, in particular, Mark Dusheiko and Neil Söderlund for their contributions to this work. Hannah Duckmanton provided excellent secretarial assistance. The author is grateful for financial support from the Department of Health although the views expressed in this paper are those of the author and do not necessarily reflect those of the Department of Health.

FURTHER COPIES

Further copies of this document (at price £7.50 to cover the cost of publication, postage and packing) are available from:

The Publications Office
Centre for Health Economics
University of York
York YO10 5DD

Please make cheques payable to 'The University of York'. Details of other papers can be obtained from the same address or telephone (01904) 433648.

SUMMARY

The 1997 White Paper, *The New NHS*, announced that Trusts would be required to publish and benchmark the costs of the treatments they provide on a similar basis. Between November 1998 and March 1999 five different indices were produced purporting to measure the unit costs of service provision in English acute Trusts. The publication of multiple indices has meant that Trusts appearing 'inefficient' on one index can emerge as relatively 'efficient' on another. This has the potential to create confusion.

Of the five indices, the reference cost index (RCI) is the least satisfactory. This index fails to fully capture the range of service provision or to account for factors known to influence hospital costs, some of which are beyond immediate managerial control. Hence, interpretation of the RCI as evidence of poor performance cannot be justified. By explicitly taking such factors into account, the casemix cost indices (CCIs) are a considerable improvement on the RCI. However, the efficiency scores generated by the CCIs are point estimates that are not accompanied by any range of uncertainty.

By replicating the econometric models used to produce the CCIs it is possible to calculate confidence intervals around the efficiency scores. Comparison of the confidence intervals suggests that differences in Trust unit costs (and, by implication, efficiency) are not statistically significant. As such, differential performance targets based on unit cost data would be unfair.

The paper discusses possible reasons why the econometric analysis fails to identify variations in unit costs. It is suggested that there may be more fruitful means to encourage performance improvements among NHS acute Trusts than general target setting based on unit cost data.

I INTRODUCTION

In its drive for greater accountability and efficiency in the National Health Service (NHS), the Labour administration has published a new set of performance indicators to measure the activity and outcomes of the health service. These replace the Health Service Indicators introduced in the 1980s. The new measures include a set of clinical indicators, which focus on the quality of care provision, and a set of cost indices, which attempt to shed light on the cost of service provision.

The cost indices are designed to replace the discredited purchaser efficiency index, which failed to capture fully the diverse activities of the health service and introduced incentives which were contrary to efficient service delivery (Appleby 1996). The government's original intention was to produce comparative information allowing Trusts, Health Authorities and Primary Care Groups to consider the costs of local service provision in the light of national data (NHSE 1997). Evidence of higher costs was to be interpreted as indicating poor performance, necessitating remedial action.

The original schedule, the Reference Cost Index (RCI), was limited in scope and failed to account for many of the factors influencing hospital costs which are outside management control. As such, it could not be inferred that those Trusts appearing to have high costs on the RCI were relatively inefficient. To rectify the limitations of the RCI, the Audit Commission and DoH produced three new schedules, the Casemix Cost Indices (CCIs). These indices attempt to take into account a range of factors deemed influential in explaining differences in unit costs observed among Trusts.

However, the efficiency scores generated by the CCIs are point estimates that are not accompanied by any range of uncertainty. Without this we cannot be confident that the apparent differences in Trust unit costs reflect true differences. Consequently, we cannot be sure that targets set on the basis of the efficiency scores are fair. This deficiency contrasts with the publication of the six clinical indicators, which aim to provide an indication of the quality of service provision. These indicators are reported with confidence intervals around each Trust's performance score, allowing users to judge whether variations among Trusts in the mean scores are truly different or merely a matter of chance (NHSE 1999a). Where confidence intervals overlap Trusts cannot be considered to differ in performance.

This paper describes the various cost indices and the interpretations that may be drawn from consulting them. In Section II the construction of the five indices is briefly described, while Section III reports how Trusts appear to vary in their relative performance across indices. Section IV describes a method to calculate confidence intervals around the efficiency scores derived from the CCIs, presents the results of these calculations and assesses whether differences in the point estimates reported for individual Trusts are true differences or merely a matter of chance. Section V outlines changes in coding behaviour that may be observed as a consequence of publication of the cost data. The final section discusses the policy implications of the analysis, concluding that it may not be fair to set unit cost targets for Trusts based on the currently available data.

II THE DEVELOPMENT OF THE NHS COST INDICES

Between November 1998 and March 1999 five different indices were produced purporting to measure the unit costs of service provision in English acute Trusts. A brief description of each index appears in Box 1.

Box 1: The NHS cost indices

Reference Cost Indices

- **RCI**: The National Reference Cost Index is compiled using data provided by Trusts about their unit costs by Healthcare Resource Group (HRG) for their main surgical specialties. The RCI is a weighted average of all HRG costs in each Trust relative to the national average. The Market Forces Factor (MFF) is then added to account for differences in local factor costs. The index was published in November 1998 and covers around 50% of Trust expenditure.
- **RCI+**: The RCI covers only acute activity. The RCI+ is more extensive and covers expenditure and activity in general, acute and maternity specialties and accident and emergency (A&E). The RCI+ covers around 70% of expenditure.

Casemix Cost Indices

- **CCI.** The Casemix Cost Index (CCI), unlike the RCI, includes mental health services and day care costs. The index is a ratio of actual to expected costs, taking into account hospital casemix. Activity in the CCI is summarised as a weighted combination of HRG based inpatient spells, outpatient first attendances and A&E first attendances.
- **2CCI**. The Casemix Costliness Cost Index (2CCI) builds on the CCI, incorporating adjustment for other variables hypothesised to explain cost differences among trusts. These variables included hospital transfers, multi-episode spells, and the proportion of elderly or female patients, student numbers, research revenue and the MFF. The extent of the adjustment for each of these variables is estimated through regression analysis.
- **3CCI.** The 3CCI – the Casemix Costliness and Configuration Index – attempts to take into account differences in hospital configuration, over and above the adjustments made in the 2CCI. These include the costs of multi-site working, hospital size, and capacity utilisation.

The five indices fall into two groups. The *reference cost indices* (RCI and RCI+) measure inpatient activity using Finished Consultant Episodes (FCEs) and are based on non-routine data provided by Trusts to the NHS Executive (NHSE) in 1998. The Department of Health (DoH) led development of the *casemix cost indices* (CCI, 2CCI and 3CCI). These indices measure inpatient activity using hospital spells, are based on routine activity and financial data, and include adjustment for a range of factors deemed influential in explaining variations in unit costs among Trusts. The construction of each index is described in more detail below.

The reference cost indices

The 1997 White Paper, *The New NHS*, announced that Trusts would be required to publish and benchmark the costs of the treatments they provide on a similar basis (NHSE 1997). The intention was to develop a national schedule of reference costs, based on allocation of accounting costs by Healthcare Resource Groups (HRGs). The first version of reference costs was published in November 1998 as the National Schedule of Reference Costs (NRSC). Trusts have long been encouraged to generate such information but this was the first time that national data had been made publicly available.[1]

As well as providing a list of HRG costs, the data used to create the NSRC were aggregated to provide a summary of each Trust's overall costs, reported as the Reference Cost Index (RCI) (NHSE, 1998). The RCI is a weighted summary of all HRG costs in each Trust relative to the national average and is adjusted for the differences in the cost of land, building and labour in the NHS using the Market Forces Factor (MFF).

The RCI initially generated publicity in the national press but little comment in health service publications, perhaps implying that it was not viewed as credible within the NHS. The fact that, even though it purported to adjust for case-mix, the two Trusts providing the most and least complicated procedures were respectively least and most 'efficient' according to the index, suggests that factors explaining cost differences were not adequately dealt with.

Criticism focused on three major deficiencies. First, the RCI related to surgical activity and failed to account for outpatient, accident and emergency (A&E), and non-acute activity. Second, it used FCEs to measure activity, a measure with known deficiencies, in view of the amount of local discretion over what constitutes the time spent in the care of a consultant (Clarke and McKee 1992). Third, although there was an attempt at clustering similar hospitals into family group, the index itself failed to take into account factors known to influence costs such as the severity of cases treated and hospital configuration.

[1] Only two Regional Offices, North West and Trent, had previously compiled cost information for Trusts in their region.

The problem of limited coverage was partially addressed by construction of a new cost index referred to as the RCI+. It used available, but incomplete, data on activity not covered by the RCI in an attempt to estimate relative cost performance over the entire range of an acute hospital's activity. However, few details have emerged about how the RCI+ was constructed and even the index itself has not been made publicly available. Nevertheless, the NHSE recommended that the RCI+ provide the starting point in regions for discussions leading to the setting of the 1999/2000 efficiency targets for individual Trusts.

The casemix cost indices

More systematic attempts to deal with the problems of the original RCI appear in the three cost indices produced by the DoH and the Audit Commission in early 1999, these being:

- CCI - Casemix Cost Index;
- 2CCI - Casemix Costliness Index;
- 3CCI - Casemix Costliness and Configuration Index.

All three new indices use patient 'spells' rather than FCEs as a measure of hospital activity, with spells corresponding to hospital admissions. Furthermore, they are based on data from the Hospital Episode Statistics rather than the non-routine returns used for the RCI. The three new indices were derived by calculating an index of expected to actual costs and then regressing this index against a succession of explanatory factors. The 3CCI includes the most comprehensive set of adjustments.

Full details about the construction of the cost indices are provided in Söderlund and van der Merwe (1999), but a brief overview is presented here. The casemix cost index (CCI) for hospital *i* is an index of actual over expected casemix weighted costs:

$$CCI_i = \frac{C_i}{[(IC*IP_i*H_i) / IP*H] + [\Sigma_j OP_{ij}*(OC_j/OP_j)] + [AE_i*AC/AE]} \quad (1)$$

where C_i is the cost of inpatient, outpatient and A&E care in hospital *i*; IC is the total cost of inpatient spells for all acute hospitals; IP_i is the number of inpatient (including day case) spells in hospital *i*; H_i is the HRG casemix index for hospital *i*; IP is the total number of inpatient spells for all acute hospitals; H is the average casemix index for all acute hospitals; OP_{ij} is the number of first outpatient attendances across all specialties in hospital *i*; OC_j is the total cost of outpatient attendances for all acute hospitals in specialty *j*; OP_j is the number of first outpatient attendances for all acute hospitals in specialty *j*; AE_i is the number of first

A&E attendances in hospital *i*; AC is the total cost of A&E attendances in all acute hospitals; and AE is the number of first A&E attendances in all acute hospitals. As well as being a Trust-level index, the CCI was also calculated at specialty level.

The Casemix Costliness Cost Index (2CCI) builds on the CCI, incorporating adjustment for other variables hypothesised to explain cost differences among Trusts. These variables are listed in the top half of Box 2. Many of these adjustments attempt to account for the possibility that, even within an HRG, some hospitals will treat more costly patients. Hospital transfers, multi-episode spells, and the proportion of elderly or female patients are included to account for cost differences over and above the HRG casemix adjustment. In addition, the 2CCI makes allowance for possible cross-subsidisation between patient care and teaching or research which may not be adequately dealt with in the funding allocations, and for differences in local factor costs, assessed using the MFF. The extent of the adjustment for each of these variables is estimated using a linear regression model of the following form:

$$CCI_i = \alpha_i + \beta \mathbf{g}_i + u_i \qquad (2)$$

where CCI_i is the average cost per casemix weighted patient in hospital *i* and **g** represents a vector of the explanatory variables. The residuals, u_i, are distributed symmetrically around the cost function with a mean of zero, and it is these which provide the 2CCI efficiency scores (υ_i), calculated as $\upsilon_i = u_i + 1$.

Box 2: Explanatory variables used in the construction of the 2CCI and 3CCI

2CCI: g	• The proportion of spells transferred into the Trust • The proportion of spells transferred from the Trust • The proportion of inter-specialty transfers • Emergency admissions as a proportion of total spells • Finished Consultant Episodes per spell • Outpatient re-attendances per spell • An index of unexpected emergency admissions • An index of within-HRG casemix complexity • The proportion of inpatients under 15 years • The proportion of inpatients over 60 years • The proportion of female inpatients • Student whole time equivalents per spell • The percentage of total income from research • The Market Forces Factor
3CCI: h	• The number of inpatient spells • The number of first outpatient attendances • The number of first A&E attendances • Average number of beds • Heated building volume per bed • The number of sites with more than 50 beds • An index of specialisation

In similar exercises in which a cost frontier is derived (eg Feldstein 1967), it is assumed that $u_i \geq 0$ for all *i*, so that hospitals lie on or above the deterministic cost function. The DoH's approach is not to derive a frontier as such, but instead to select a hospital to serve as a reference point against which the efficiency of the remainder can be compared. The published indices are centred around 1, and the average hospital provides the reference case. The two methods produce identical rankings of hospitals but, unlike the frontier method, the DoH approach does not imply that any hospital is 100% efficient.

The 2CCI is described as a long-run cost index in that it implies that hospital re-configuration is feasible in the long term. In reality, hospitals may appear relatively inefficient on the 2CCI because of factors beyond immediate managerial control. The 3CCI attempts to take some of these factors into account, over and above the adjustments made in the 2CCI. These include the costs of multi-site working, measured by the number of sites with more than 50 beds; hospital size, measured by the number of beds; and capacity utilisation, reflected by the number of patients treated. The model takes the following form:

$$CCI_i = \alpha_2 + \beta \mathbf{g}_i + \lambda \mathbf{h}_i + v_i \qquad (3)$$

where **h** is a vector of the additional explanatory variables relating to hospital configuration. The short run efficiency scores produced for each hospital v_i are calculated as $v_i = v_i + 1$.

III COMPARISONS ACROSS INDICES

The RCI was intended to summarise comparative cost information so that NHS providers and commissioners would be able to identify poor performance and engage in informed debate about opportunities for improvement. Where the RCI has identified supposedly high cost providers, these Trusts have come under pressure from their local Health Authorities to explain their position and to improve their performance. Regional targets for efficiency improvements have also been set with reference to the RCI and RCI+. However, the reference cost indices fail to fully capture the range of service provision or to account for factors known to influence hospital costs, some of which are beyond immediate managerial control. Hence, interpreting these indices as evidence of poor performance may be misleading, especially in the absence of other information about cost influencing factors.

The more recent indices have been developed in an attempt to rectify the shortcomings of the RCI and RCI+. However, the publication of multiple indices has meant that Trusts appearing 'inefficient' on one index can emerge as relatively 'efficient' on another. This has the potential to create confusion.

To some extent, changes from one index to the next in how Trusts appear to perform relative to their peers is to be expected, as Trust 'league positions' (or ranks) reflect the differential impact of the various explanatory factors that each index takes into account. The DoH and Audit Commission provided advice to guide local interpretation when the CCIs were published, pointing some of the reasons for potential discrepancies between the RCI and 2CCI (Audit Commission and DoH, 1999).

However, while these adjustments will lead to some change in ranks, what is surprising is the extent of movement experienced by individual Trusts. Of the 213 Trusts for which full data are available, only ten remain in the same decile across all indices. In contrast, Trusts move an average of 80 places across indices – equivalent to a third of the 'league'. At the extreme, The Queen Victoria Hospital Trust moves from near the bottom of the RCI (206/213) to third most 'efficient' on the 3CCI.

This variability in league positions ranks across indices can be represented by 'Adjustment Intervals'.[2] The figures in Appendix A shows the Adjustment Interval for each acute Trust, categorised by family group. The scale for each figure ranges from 1 to 213, with 1 indicating the lowest cost provider according to each index. The lines indicate the range in ranks for each Trust across the RCI, CCI, 2CCI and 3CCI (data for the RCI+ are unavailable). The length of the line reflects the impact on each Trust's league position of the adjustments used in the construction of each index, such as the move from FCEs to spells or the influence of the various explanatory variables incorporated in the 2CCI and 3CCI. Short lines indicate that the adjustments have little effect on the Trust's relative position, such as with for the George Eliot Hospital (figure A1), which appears to have relatively low costs whichever index is used. Trusts with longer lines are more sensitive to the adjustment process, as is the case for Bedford Hospital and The Royal West Sussex (figure A1).

In most family groups the lines overlap to a large extent. This demonstrates the danger of using the RCI as the single measure of Trust unit costs. The adjustments incorporated in the construction of the CCIs have a marked effect on the relative performance of individual

[2] This follows the terminology adopted in a paper by Goldstein and Spiegelhalter (1996) discussing statistical issues in institutional performance, in which it was recommended that confidence intervals be constructed around ranks to capture uncertainty, and that adjustment procedures be conducted on the raw scores (in this case, unit costs).

Trusts. If Trusts are under pressure from Health Authorities to justify 'poor performance' according to the RCI, they would probably benefit from checking how much their position changes when using another index before tackling the apparent 'problem'.

What are the implications of this apparent variability? First, it is important to be clear about which index is the most appropriate measure of unit costs. There are good economic reasons for believing that the 3CCI should be a much better rough indicator of short-term relative performance than the other indices because it better isolates the things over which management may have influence. Management has little control over size, mergers reflect national policy, diseconomies of scale should not be confused with managerial 'slack' and adjustment of the capital stock associated with multi-site working takes time and depends on regional capital priorities. By attempting to control for these factors, the 3CCI is more likely to reflect the performance of the Trust rather than the environment in which it operates.

Second, it is important that each index reports not only each Trust's relative position, but also provides an indication of the confidence that might be placed on this position being an accurate reflection of relative efficiency. It is well known that crude rankings can lead to erroneous conclusions about relative performance (Goldstein and Spiegelhalter 1996). It may be that there is considerable movement across indices because the measured performance of different Trusts does not actually vary greatly but is estimated with a relatively large error term. Overlapping confidence intervals would be observed if Trusts have broadly similar performance. If so, this would suggest that variations in efficiency among Trusts are not great enough to merit their use for setting differential efficiency targets. The next section addresses this possibility.

IV COMPARISONS WITHIN INDICES

While the CCIs are an improvement on the RCI, the published efficiency scores are point estimates that are not accompanied by any range of uncertainty. In order to judge whether the efficiency scores represent true differences in performance it is necessary to calculate confidence intervals around each Trust's score.

As described in Section II, the 2CCI and 3CCI efficiency scores (υ_i or ν_i) for each Trust are derived from the residuals from a least squares regression, performed on non-hierarchical cross-sectional data.[3] Although rarely calculated, it is possible to compute the variance

[3] Different techniques are available for calculating confidence intervals when data are hierarchical or when there are repeat observations over time. For a discussion of these techniques see Goldstein and Speigelhalter (1996) and Marshall and Speigelhalter (1998).

around the residual for each Trust (Maddala 1988, p409-11). Confidence intervals can be calculated once the variance is known. Consider a model with no constant term:

$$\mathbf{y} = \mathbf{x}_i\mathbf{b} + \varepsilon \tag{4}$$

where $\mathbf{y}$ is the vector of observations on the dependent variable; $\mathbf{x}_i$ is a vector of explanatory variables; $\mathbf{b}$ is a vector of unknown parameters; and ε is a vector of unobservable errors, with zero mean and constant variance.

The model yields predictions ($\hat{\mathbf{y}}$) from the least squares regression

$$\hat{\mathbf{y}} = \mathbf{x}_i\hat{\mathbf{b}} \tag{5}$$

where $\hat{\mathbf{b}}$ is a vector of estimated coefficients. $\hat{\mathbf{y}}$ has a variance due to the variance of the estimated coefficients. This reflects the error in our prediction due to our uncertainty about what the parameter estimates ($\mathbf{b}$) actually are. This uncertainty can be expressed as:

$$\mathrm{var}(\hat{\mathbf{y}}) = \mathrm{var}(\mathbf{x}_i\hat{\mathbf{b}}) = s^2 h_i \tag{6}$$

where s^2 is the mean square error of the regression and $h_i = \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T$, where $\mathbf{x}_i$ is a vector of k independent variables for the i^{th} observation, $\mathbf{X}$ is a n x k matrix of observations on the independent variable, $\mathbf{X}^T$ is the transpose of this matrix, and $(\mathbf{X}^T\mathbf{X})^{-1}$ is the inverse of their product. h_i is a diagonal element of the projection matrix and is commonly termed the 'hat matrix'. As an example, in the single variable case, this gives an indication of the proximity of the i^{th} observation to the mean of the other observations divided by the sum of the deviation for all observations from the mean.

From (6) we can derive the standard error of the prediction:

$$s_{p_i} = s\sqrt{h_i} \tag{7}$$

Although the errors (ε) are unobservable, the model yields residuals, ($\mathbf{e}$), such that $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$. Moreover, $\mathrm{var}(\mathbf{y}) = \mathrm{var}(\hat{\mathbf{y}}) + \mathrm{var}(\mathbf{e}) = s^2 h_i + s^2$. We can decompose s^2 into prediction and residual variances:

$$s^2 = s^2 h_i + s^2(1 - h_i) \tag{8}$$

where $s^2(1\text{-}h_i)$ is the residual variance. Thus the standard error of the residual is:

$$s_{r_i} = s\sqrt{(1 - h_i)} \tag{9}$$

Details of the residuals and their associated variances are summarised by family group for the 2CCI in Table 1 and for the 3CCI in Table 2. These data were produced after re-estimating equations (2) and (3) using data for 1995/96, these data having been provided on the NHS web site [http://tap.ccta.gov.uk/doh/trustben.nsf]. As for the original estimation process, the re-analysis excludes Trusts identified as outliers.[4]

The efficiency scores for the 2CCI and 3CCI are derived from the residuals (u_i and v_i respectively) which are centred at zero, this being the value for the average Trust. Trusts with residual values less than zero are interpreted as being more efficient (ie having lower unit costs) than average. The published efficiency scores are simply scaled up so that the average Trust has a value of 1.

Consideration of the mean value of the residuals produced by the 2CCI for each family group implies that specialist (code 140), teaching (210) and multi-service (310 and 313) Trusts have generally higher unit costs than other hospitals: the mean of u_i for these family groups is greater than zero (Table 1). This at least accords with the theoretical basis of the 2CCI and supports the use of the 3CCI, in which the scale of service provision and the configuration of the institutions is formally taken into account. Once these factors are considered, these groups of Trusts (with the exception of acute teaching Trusts) appear to have relatively lower unit costs than was implied by the 2CCI, with their mean residual values (v_i) moving closer to zero (Table 2). The 3CCI also has a tighter distribution of residuals, the overall standard deviation being 0.081 compared to 0.105 for the 2CCI.

Even casual consideration of the range of values for u_i and v_i suggests that users should be wary about attributing statistical significance to the differences in mean residuals for the family groups. The range between the Trusts with the lowest (min) and highest (max) residuals within family groups overlaps across all families.

Summary statistics relating to the variance around the residuals, var(u_i) and var(v_i), are also shown in Tables 1 and 2. It is worth noting the seemingly counter-intuitive estimated range of uncertainty around the efficiency scores for some Trusts. It might be expected that we are more uncertain about the efficiency scores of (say) specialist Trusts than we are for other hospitals, perhaps because HRGs fail to adequate capture their more specialised casemix. However, the data do not reflect this, with the standard deviation around the 3CCI residuals (Table 2) for specialist hospitals (0.070) being less than for all hospitals (0.081). The reason

[4] Outliers were identified using the DFITS procedure, and applying a cut-off point where DFITS>3*$(k/n)^{0.5}$ where k is the number of parameters estimated and n the number of observations (Söderlund and van der Merwe, 1999).

is that the efficiency score is derived from the residuals, and the uncertainty around the residuals stems directly from the relationship between $\text{var}(\hat{y}_i)$ and $\text{var}(\hat{\varepsilon}_i)$. Since $\text{var}(\hat{y}_i) = h_j s^2$ and $\text{var}(e_i) = (1- h_j)s^2$, predicted values for observations further from the regression line will have relatively large variances and their corresponding residuals will have relatively small variances (Cook and Weisberg 1982, p14).

Table 1: Residuals (u_i) and variance (Var(u_i)) from the 2CCI, by family group

2CCI		u_i				Var(u_i)			
Family Code	*n*	Mean	SD	Min	Max	Mean	SD	Min	Max
111	29	-0.016	0.085	-0.159	0.142	0.107	0.001	0.101	0.108
121	44	-0.029	0.103	-0.254	0.222	0.107	0.001	0.101	0.108
131	41	-0.022	0.081	-0.201	0.252	0.106	0.004	0.083	0.108
135	19	-0.013	0.083	-0.181	0.176	0.104	0.008	0.073	0.107
140	16	0.013	0.115	-0.167	0.275	0.091	0.011	0.056	0.104
210	27	0.006	0.130	-0.189	0.345	0.102	0.006	0.083	0.107
310	25	0.033	0.106	-0.153	0.224	0.106	0.004	0.088	0.108
313	16	0.098	0.100	-0.032	0.270	0.107	0.001	0.105	0.108
All trusts	*217*	*0.000*	*0.105*	*-0.254*	*0.345*	*0.104*	*0.006*	*0.056*	*0.108*

Family codes: 111 Small/medium acute Trusts, outside London; 121 Large acute Trusts, outside London; 131 Very large acute Trusts, outside London; 135 Acute Trusts, London; 140 Specialist Trusts; 210 Acute teaching Trusts; 310 Small/medium/large multi-service Trusts; 313 Very large multi-service Trusts

Table 2: Residuals (v_i) and variance (Var(v_i)) from the 3CCI, by family group

3CCI		v_i				Var(v_i)			
Family Code	*n*	Mean	SD	Min	Max	Mean	SD	Min	Max
111	29	-0.014	0.068	-0.129	0.154	0.083	0.002	0.076	0.084
121	44	-0.028	0.082	-0.257	0.149	0.083	0.001	0.078	0.085
131	41	0.015	0.062	-0.104	0.185	0.082	0.004	0.062	0.084
135	19	0.015	0.073	-0.095	0.191	0.081	0.007	0.055	0.084
140	16	0.006	0.104	-0.088	0.304	0.070	0.009	0.043	0.080
210	27	0.008	0.086	-0.113	0.220	0.077	0.005	0.062	0.082
310	25	0.003	0.099	-0.180	0.182	0.082	0.004	0.067	0.084
313	16	0.016	0.081	-0.090	0.164	0.083	0.001	0.080	0.084
All trusts	*217*	*0.000*	*0.081*	*-0.257*	*0.304*	*0.081*	*0.006*	*0.043*	*0.085*

Family codes: 111 Small/medium acute Trusts, outside London; 121 Large acute Trusts, outside London; 131 Very large acute Trusts, outside London; 135 Acute Trusts, London; 140 Specialist Trusts; 210 Acute teaching Trusts; 310 Small/medium/large multi-service Trusts; 313 Very large multi-service Trusts

Figure 1 presents the 3CCI point estimates of the efficiency scores and confidence intervals surrounding these estimates for all 217 Trusts, ordered according to the estimated efficiency score. These data are provided in Appendix B for each Trust, grouped by family group, in a form similar to the figures presented for the clinical indicators (NHSE 1999a). Trusts that have merged since the 1995/96 are named in capital letters. The scale is centred at 1, this indicating the mean efficiency score for all English NHS acute Trusts.

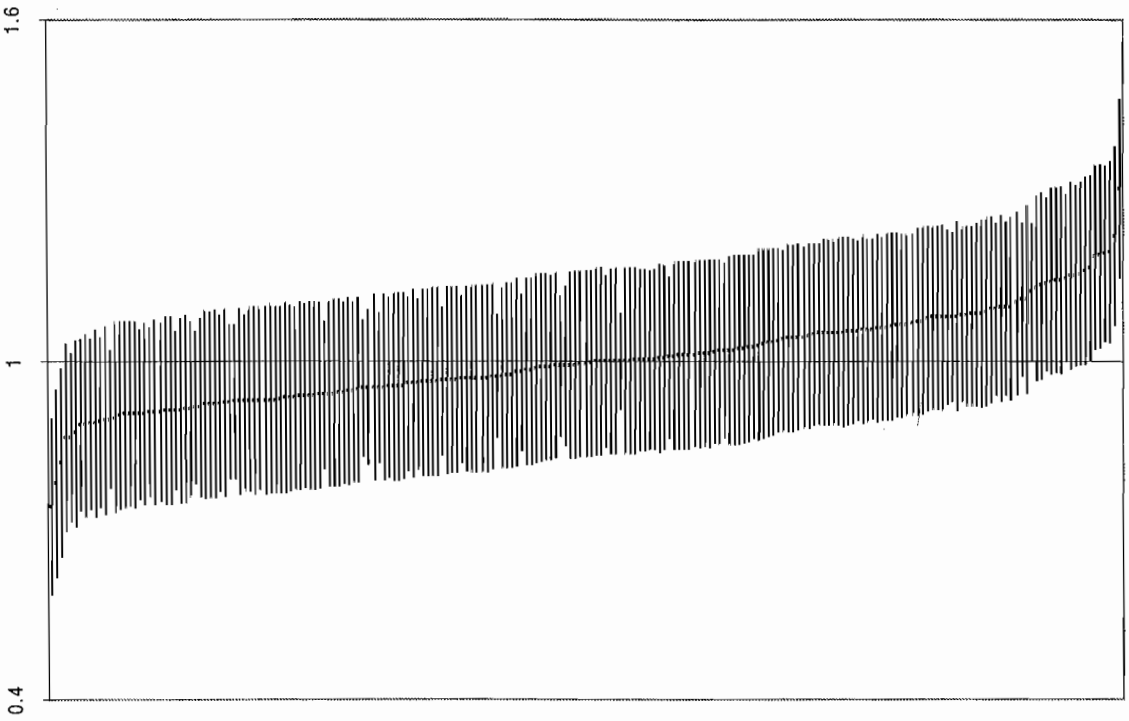

Figure 1: Pairwise 95% confidence intervals based on 3CCI residuals

Confidence intervals around the efficiency scores of individual Trusts are calculated as $e_i \pm 1.96 \times s_{r_{ij}}$. The use of 1.96 to mark the boundaries of the confidence intervals is consistent with the approach adopted by the Department of Health in reporting 95% confidence intervals around the high level performance intervals for Trusts and health authorities (NHSE 1999a). This is appropriate when making pairwise comparisons. However, to maintain a 95% level of significance when making multiple comparisons we need to adjust the critical level to account for the number of Trusts being compared.

In order to make multiple comparisons, the critical level has been adjusted using the Šidák adjustment (Šidák 1967) which preserves a 95% significance level. The confidence interval when making comparisons across all Trusts is now calculated as $e_i \pm 3.68 \times s_{r_{ij}}$, and figure 2 presents the Trust series with these revised confidence intervals.

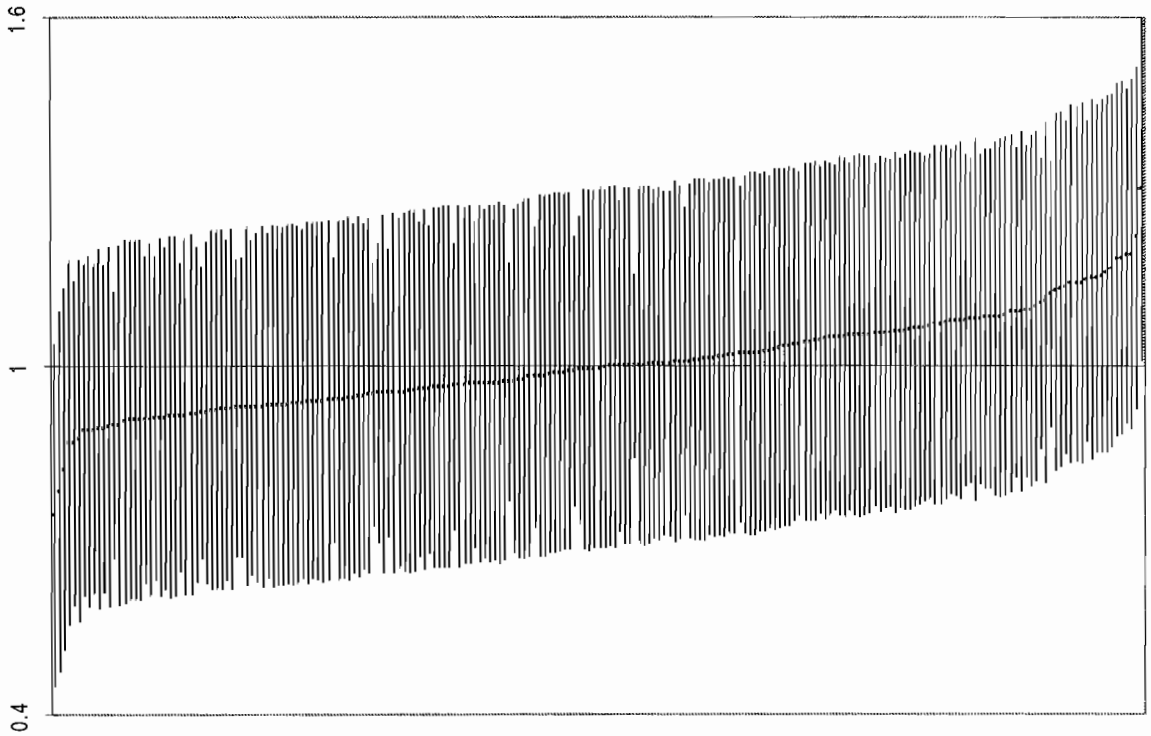

Figure 2: Multiple comparison 95% confidence intervals based on 3CCI residuals

The critical values applied for comparisons within family groups are shown in Box 3. These adjustments suggest that no significance can be attached to the mean differences in unit costs observed among Trusts. In other words, we cannot confidently claim that the Trusts that appear among the most efficient under the 3CCI are actually any more efficient than those Trusts that appear among the least efficient.

Box 3: Significant differences within family groups in 3CCI unit costs

Family code	*Number*	*Critical value*	*Significant differences among Trusts*
All Trusts	217	± 3.68	No significant differences.
111	29	± 3.13	No significant differences.
121	44	± 3.25	No significant differences.
131	41	± 3.23	No significant differences.
135	19	± 3.00	No significant differences.
140	16	± 2.95	No significant differences.
210	27	± 3.11	No significant differences.
310	25	± 3.09	No significant differences.
313	16	± 2.95	No significant differences.
Family codes: 111 Small/medium acute Trusts, outside London; 121 Large acute Trusts, outside London; 131 Very large acute Trusts, outside London; 135 Acute Trusts, London; 140 Specialist Trusts; 210 Acute teaching Trusts; 310 Small/medium/large multi-service Trusts; 313 Very large multi-service Trusts			

V THE IMPACT ON DEFINING COSTS AND ACTIVITY

In early 1999, the Audit Commission and DoH published Trust profiles based on the CCIs for all acute providers in England. Each acute Trust was issued with an explanatory booklet (Audit Commission and DoH, 1999) and a profile tailored to the individual Trust which:

- Categorised the Trust into one of eight family groups, on the basis of size, location and teaching status;
- Set out a graphical comparison of their Trust level unit costs, expressed by the RCI, 2CCI and 3CCI, relative to their family group peers;
- Presented a graphical comparison of their specialty level unit costs, based on the CCI, relative to those of their family group peers which have similar specialties;
- Provided additional graphical information considered relevant to the analysis of cost differences, including average length of stay, casemix complexity, day case activity and outpatient costs.

The profiles were sent to the finance directorate at each Trust, and Regional Offices received a copy for all Trusts in their region. They were not circulated more widely, the intention (presumably) being that they are used to inform internal discussions in the first instance without Trusts having to defend their position to their local Health Authorities and Primary Care Groups.[5]

The advantages of restricting circulation are that Trusts:

- had time to familiarise themselves with the techniques;
- could supplement the information provided in the profiles with their own data, and initiate internal management processes to identify areas warranting improvement;
- assess whether their costing and counting practices differed from other Trusts, particularly as the profiles present data at specialty level.

Any system designed to measure hospital performance is at the mercy of the data underpinning it. Also, the publication of performance data can be expected to influence future data collection. This section considers how the Trust profiles might affect the production of cost and activity data as a result of Trusts receiving the profiles.

[5] Although the indices and the data used in their construction were made available on the NHS Web [http://tap.ccta.gov.uk/doh/trustben.nsf].

Re-allocating costs

Teaching costs

The data used to calculate the cost indices (ie C_i in equation 1) are based on expenditure attributable to the delivery of inpatient, outpatient and A&E expenditure. Income from private patients and that relating to research and undergraduate teaching is excluded. There are two problems with the exclusion of teaching activities, and these imply that Trusts are not being compared on an equivalent basis.

First, the revenue Trusts receive for providing undergraduate teaching is probably an inaccurate reflection of the costs incurred. Funds for teaching are allocated to Regions using the SIFT formula, and Regions allocate funds to their teaching Trusts mainly on the basis of facilities with the number of student placements having less weight. Although the research element of the old SIFTR formula has been reviewed recently (Culyer 1994), the teaching component continues to rely on information on the excess costs per student dating from 1984/85 and is known to subsidise historic infrastructure costs (Bevan 1999). Those teaching Trusts that are over-compensated for their undergraduate teaching costs will be able to report lower costs of service provision than those of their peers that are under-compensated.

The second problem is that teaching Trusts enjoy considerable discretion about what to categorise as teaching expenditure and this may disguise the true cost of service provision. The problem stems, in large part, from an inability to determine accurately the excess costs of teaching over and above service provision. These excess costs may arise from longer theatre sessions or from keeping patients in hospital beyond 'normal' requirements so that students can see them, although the extent to which these are influenced by the presence of medical students is unknown. The uncertainty allows Trusts a reasonable amount of discretion about how to account for their expenditure. In one Trust it is believed possible to categorise up to £1m to either teaching or patient services, representing almost 1% of the Trust's total income. Clearly, such flexibility will have a substantial effect on unit costs.

The 2CCI and 3CCI attempt to identify cross-subsidisation over and above the SIFT allocations by including teaching (and research) as explanatory variables of unit costs. But there remains the problem of how to measure these activities. The adjustment for teaching activity was estimated using data on student whole time equivalents dating from 1992/93, and this adjustment suggested that SIFT failed to fully compensate for teaching costs. Subsequent work by the DoH on this issue has used more recent data on student teaching activity and no longer supports the earlier finding that teaching hospitals are under-compensated. These results have not yet been published.

Specialty allocations

Although it does not effect the Trust-level indices, the specialty-level indices based on the CCI are strongly influenced by the cost allocation procedures used in compiling the Trust Financial Return by programme and specialty (TFR2). Cost apportionment (of overheads and staff) continues to vary among providers, despite the release of updated costing guidance in January 1999 (NHS Executive 1999b).

In the past, Trusts have had little incentive to revise traditional apportionment methods or to consider practice elsewhere, simply because the data have rarely been used. Many Trusts expend little effort in compiling the TFR2, simply rolling figures forward year on year with an uplift, rather than undertaking a thorough evaluation. As Trusts do not routinely collect the information in the manner required to compile the TFR2, improving accuracy may require substantial staff input. For example, high costs in one specialty may be offset by lower costs in another, perhaps because consultants work in both specialties. This is common across obstetrics and gynaecology; urology and general surgery; and plastic surgery and dermatology. Ascertaining the appropriate apportionment of consultant costs is not a simple task.

The provision of the Audit Commission Profiles is the first instance of TFR2 data being fed back to Trusts in a standardised format. This feedback may provide the impetus for Finance Departments to reconsider their traditional apportionment methods, if only because the profiles provide comparative information. If the CCIs are to be used by Health Authorities and Primary Care Groups as 'levers for change' at specialty level, it will be more important that Trusts continue to standardise their costing practice.

Re-classifying activity

From FCEs to spells

The use of patient spells rather than FCEs in the construction of CCIs is a welcome move. FCEs have long been criticised as a measure of hospital activity, in view of differences in counting internal transfers and the use of admission wards as a means of creating the appearance of more activity.

Spells measure the entire hospital stay and may comprise more than one FCE, if patients fall under the responsibility of more than one consultant during their stay. This raises questions about which specialty patients should be allocated to when there are internal transfers among

consultants in different specialties. Should the admitting specialty or the specialty in which the patient stayed longest be used? As with cost allocation decisions, the choice will influence the specialty level CCIs, although it will not impinge on the Trust-level indices.

Of more general concern is the impact of moving to spells on casemix measurement. HRGs are based on FCEs and there are concerns about whether HRGs are valid descriptors of spell based activity, particularly when patients have been assigned to multiple HRGs during a single spell because they have been seen by more than one consultant. The National Casemix Office has conducted preliminary analysis in order to derive a single HRG for a spell (Benton *et al*, 1998). Various methods have been explored and the favoured approach is to use the most resource intensive HRG to represent the spell and ignore the other HRGs to which the patient may have been allocated. Clearly, this entails some loss of information.

Outpatient activity

There remain inconsistencies in the way Trusts distinguish outpatient, day care, day case and inpatient activity, and greater definitional guidance is required. Without clear rules, the new indices might frustrate efforts to improve efficiency by undertaking more work on an outpatient basis, because of the lower weight accorded this work in the indices. This is particularly true of the RCI, which covers only acute inpatient activity. By default, outpatient activity is accorded no value in this index. Indeed, attempts by Trusts to transfer more of their activity to outpatient settings are 'rewarded' by a poorer RCI index value, as overhead costs are spread over a smaller activity base. To overcome this perverse incentive, Trusts may respond by counting some activities (eg endoscopies) as day case procedures rather than, accurately, as outpatient activity.

This is not as problematic in the CCIs, which incorporate inpatient, outpatient and A&E activity. The effects of counting outpatient attendances as day case (ie inpatient) activity is more difficult to predict. Changing counting practices (and not re-apportioning costs) may improve the CCI score, but this depends on the extent to which the Trust's casemix complexity index (H_i in equation 1) decreases. Given its greater coverage and the less predictable effect of re-allocating activity, it can be expected that perverse incentives to deliver care in inappropriate settings will be less evident for the CCIs than for the RCI.

VI DISCUSSION

The 1998/99 financial year saw the publication of five indices designed to measure the unit costs of NHS acute Trusts in England. These indices differ in various respects, and these differences lead to very different conclusions about the relative performance of most Trusts. To avoid confusion it is important that users are aware of the construction and relative merits of each index.

Of the five indices, the RCI is the least satisfactory. It is based on *ad hoc* returns, the accuracy of which cannot be verified; it has only partial coverage of the range of hospital activity and militates against moves to offer more care on an outpatient basis; inpatient activity is measured using FCE rather than spells; and it fails to adjust for known influences on hospital costs. Despite these limitations, it appears that the reference cost indices are being used by Health Authorities as the starting point for assessing unit cost performance, if only because the CCIs have not been widely circulated. It is expected that the NHS Executive will produce a version of the RCI based on 1998/99 data in late 1999.

The CCIs are a considerable improvement on the RCI and their use should be encouraged in preference to the RCI. Trusts under pressure from their local Health Authorities to justify their apparently 'inefficient' performance on the RCI should refer to the CCIs before engaging in detailed analysis of their cost structures. It may be that the other indices present a different picture of their relative performance.

Having said this, the CCIs require further refinement. The cautious approach adopted by the Audit Commission and DoH in limiting their distribution suggests that they are unwilling to use them for target setting in the first instance. The profiles may encourage Trusts to re-evaluate their costing and counting practices, and this may iron out some of the anomalies in current practice. Improved data quality will enhance the credibility of the CCIs and, despite their greater complexity, this is likely to make them more acceptable for target setting than either the RCI or the RCI+.

However, the published efficiency scores generated by the CCIs are point estimates that are not accompanied by any range of uncertainty. Without this users cannot be confident that the supposed relative efficiency of Trusts reflects true differences. As such it is not possible to judge whether differential targets set on the basis of the efficiency scores are fair.

The analysis reported in Section IV suggests that most of the differences in the unit costs observed among English NHS Trusts are not statistically significant. The variance around the residuals in this cross-sectional analysis reflects uncertainty in the precision with which the explanatory variables have been estimated. It may be that different confidence limits could be calculated around Trust unit costs if other data are available. For example, with patient level data, it would be possible to undertake hierarchical (multi-level) analysis. Uncertainty around Trust level estimates could be estimated with reference to variance in the unit costs of individual patients within each Trust. Another approach would be to perform panel-data estimation, with repeat observations available over time for each particular Trust. However, this would require temporal consistency in the unit of analysis (or some way of dealing with Trust mergers) and in data coding and collection.

It is worth considering why there do not appear to be significant differences in unit costs among Trusts. There are two main reasons.

First, when the CCI is calculated, data on inpatient activity are trimmed to exclude outlier patients, those with 'exceptionally' long lengths of stay. The trim-point (t_{HRG}) for exclusion is HRG specific and is calculated as $t_{HRG}=Q3_{HRG} + 1.5(Q3_{HRG}-Q1_{HRG})$, where Q1 and Q3 are, respectively, the lower and upper quartile length of stay for all cases in England for the particular HRG. Trimming typically removes around 5-10% of cases but, because the outliers are those with long stays, the exclusion can result in 20-30% of bed-days being lost to the analysis (National Casemix Office 1997, p51). Trimming has the advantage of stability, in that comparisons among providers are less influenced by extreme cases. However, it may be that by excluding outlier cases from the analysis, we have removed one of the main reasons why Trust costs might differ. It may be that we can learn more about Trust efficiency by studying how exceptional cases are managed than we can from focusing on more common cases. The implication of trimming is that we may have already removed much of the difference among Trusts before undertaking any comparative analysis.

Second, to a large extent Trust income (and, therefore, by definition expenditure) is determined not by the performance of the Trust but by the NHS Resource Allocation formula. The formula attempts to allocate resources on the basis of the needs of the population. If the formula is a true reflection of population needs, the main differences in Trust income would be explained by local populations accessing services differently (either less frequently or seeking care from other sectors). Assuming that differential access is not substantial, it should come as little surprise that the unit costs of service provision do not vary significantly. After all, they are largely the outcome of a managed process of top down resource allocation.

In view of the lack of statistical significance in their unit costs, the attention accorded supposed differences in the relative efficiency of acute Trusts must be questioned. Putting pressure on Trusts to identify methods to improve their unit cost performance diverts attention from other non-quantified matters (Nutley and Smith 1998). Moreover, external organisations (Regional Offices or Health Authorities) might interpret above average unit costs as 'evidence' of poor performance and set efficiency targets on this basis. This paper has demonstrated that the apparent differences in the efficiency of individual Trusts are not statistically significant. Consequently, differential targets justified on the basis of the point estimates would be unfair.

In practice, cost improvement targets are likely to be set after consideration of a range of information, rather than merely the analysis of unit costs. The NHS reforms in which the internal market was introduced were founded on the belief that substantial inefficiency existed among NHS providers. Moreover, it was felt that efficiency would be improved by issuing general policy directives aimed at all providers. Whether further gains are to be made from this approach is arguable in the light of evidence that differential performance in terms of unit costs is not substantial.

It may be that there is more to be gained from a focus on specific Trusts or practices within Trusts. Comparisons could be undertaken on a case-study basis, allowing an attention to detail that is not possible when performing general analyses based on secondary data. Such case-studies might consider the nature of internal financial incentives (such as devolved budgeting systems), variations in organisational structures, management and accountability. While recognising that many of these factors may be difficult to assess and quantify, it could be a more fruitful approach to further improving Trust performance than the general comparative methods currently being pursued. Moreover, it would avoid setting unfair targets based on quantitative data which, on closer examination, fail to identify variations in efficiency.

REFERENCES

Appleby J. Promoting efficiency in the NHS: problems with the labour productivity index. *BMJ* 1996 313: 1319-21.

Audit Commission and Department of Health. NHS Trust Profiles Handbook – 1997/98. Audit Commission, London, 1999.

Benton PL, Azorin D, Sutch S. Linkage of patient episodes into spells. Proceedings of the National Casemix Conference, Brighton, 8-9 September 1998.

Bevan G. The medical service increment for teaching (SIFT): a £400m anachronism for inefficiency and inequity in the English NHS? *BMJ* 1999; 319: 908-911.

Clarke A, McKee M. The consultant episode: an unhelpful measure. *BMJ* 1992; 305: 1307-8.

Cook RD, Weisberg S. Residuals and influence in regression. London: Chapman and Hall, 1982.

Culyer A. Supporting research and development in the NHS. September 1994. London: HMSO.

Feldstein MS. Economic analysis for health service efficiency: econometric studies of the British National Health Service. North-Holland, Amsterdam, 1967.

Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Statist Soc* 1996; 159A (part 3): 385-443.

Maddala GS. Introduction to econometrics. London: Collier Macmillan, 1988.

Marshall EC, Spiegelhalter DJ. Comparing institutional performance using Markov chain Monte Carlo methods. In: Everitt BS, Dunn G Statistical analysis of medical data: new developments. London: Arnold, 1998.

National Casemix Office. Compendium to Version 3 HRG Documentation Set. NHS Executive G7033, volume 2, July 1997.

NHS Executive. The new NHS: modern dependable. Cm3807, Leeds, 1997.

NHS Executive. Reference costs: a consultation document. Leeds, 1998a.

NHS Executive. 1998 Reference costs. Leeds, 1998b.

NHS Executive. Quality and performance in the NHS: high level performance indicators. Leeds, 1999a.

NHS Executive. NHS costing manual. Leeds, 1999b.

Nutley S, Smith P. League tables for performance improvement in health care. *J Health Serv Res Policy* 1998 3(1) 50-57.

Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 1967 62: 626-633.

Söderlund N, van der Merwe R. Hospital benchmarking analysis and the derivation of cost indices. Centre for Health Economics Discussion paper 174, York: University of York.

APPENDICES

Appendix A: Adjustment Intervals

Note to figures. For each Trust, the following figures present the range in league positions across four unit cost indices: RCI, CCI, 2CCI and 3CCI. Although the figures are decomposed into family groups, the scale covers all English NHS acute Trusts and ranges from 1 'lowest unit cost' to 213 'highest unit cost'.

Appendix B: Confidence Intervals

Note to figures. For each Trust, the following figures present the estimated efficiency score for the 3CCI and the associated 95% confidence interval. Although the figures are decomposed into family groups, the scale covers all English NHS acute Trusts and is centred at 1, which represents the average Trust.

Figure A1 Small/Medium Acute Outside London

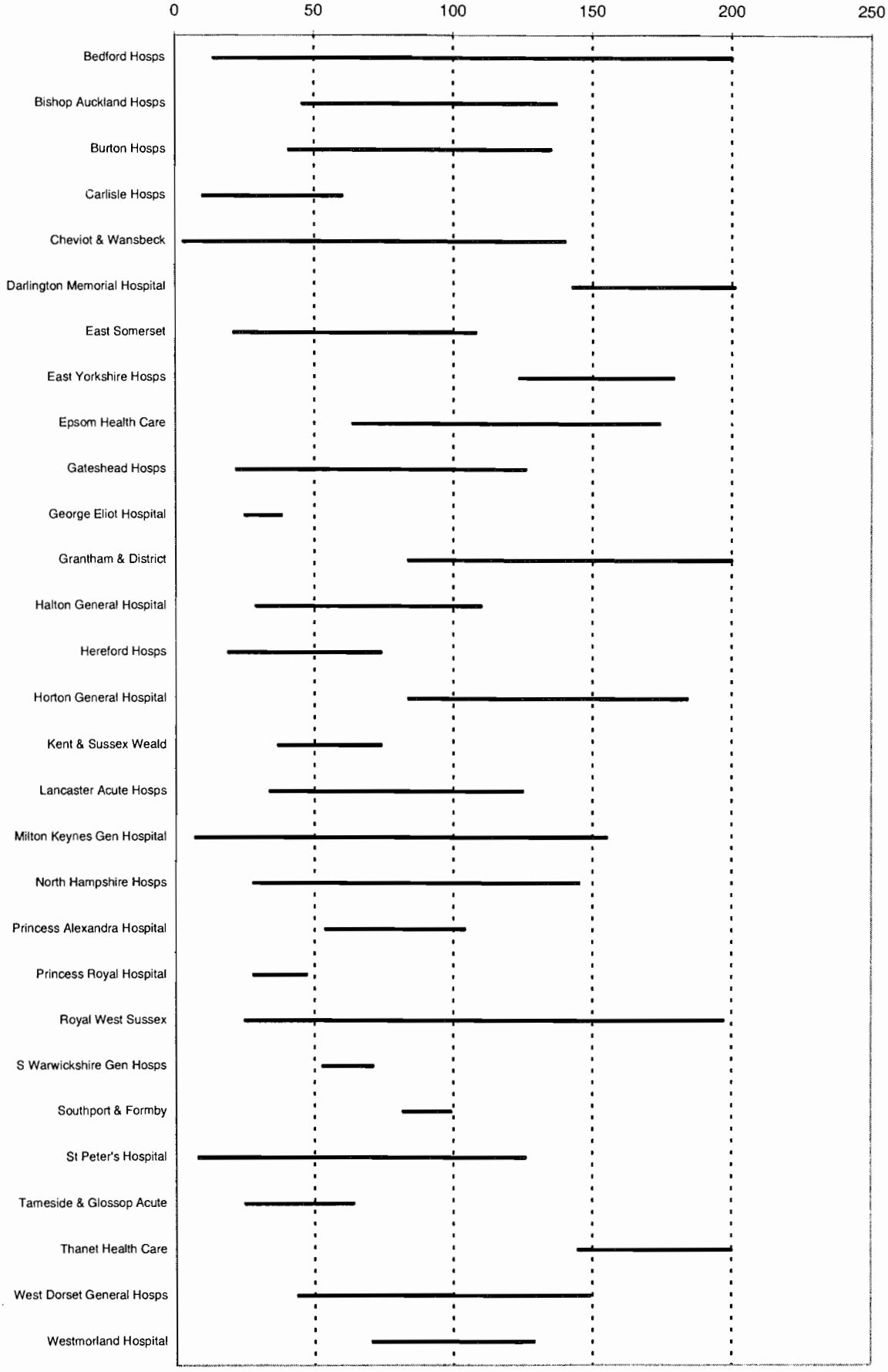

Figure A2 Large Acute Outside London

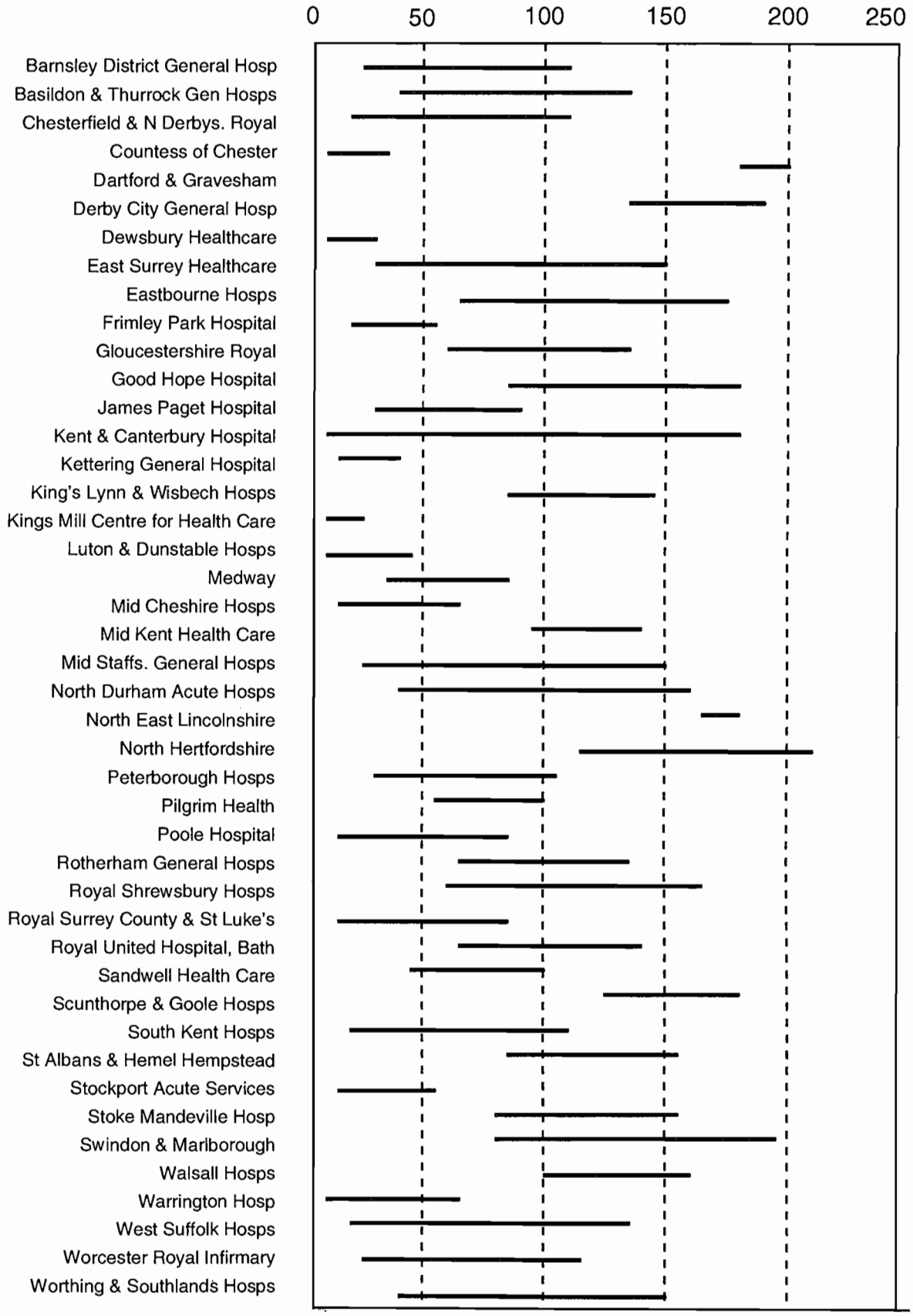

Figure A3 Very Large Acute Outside London

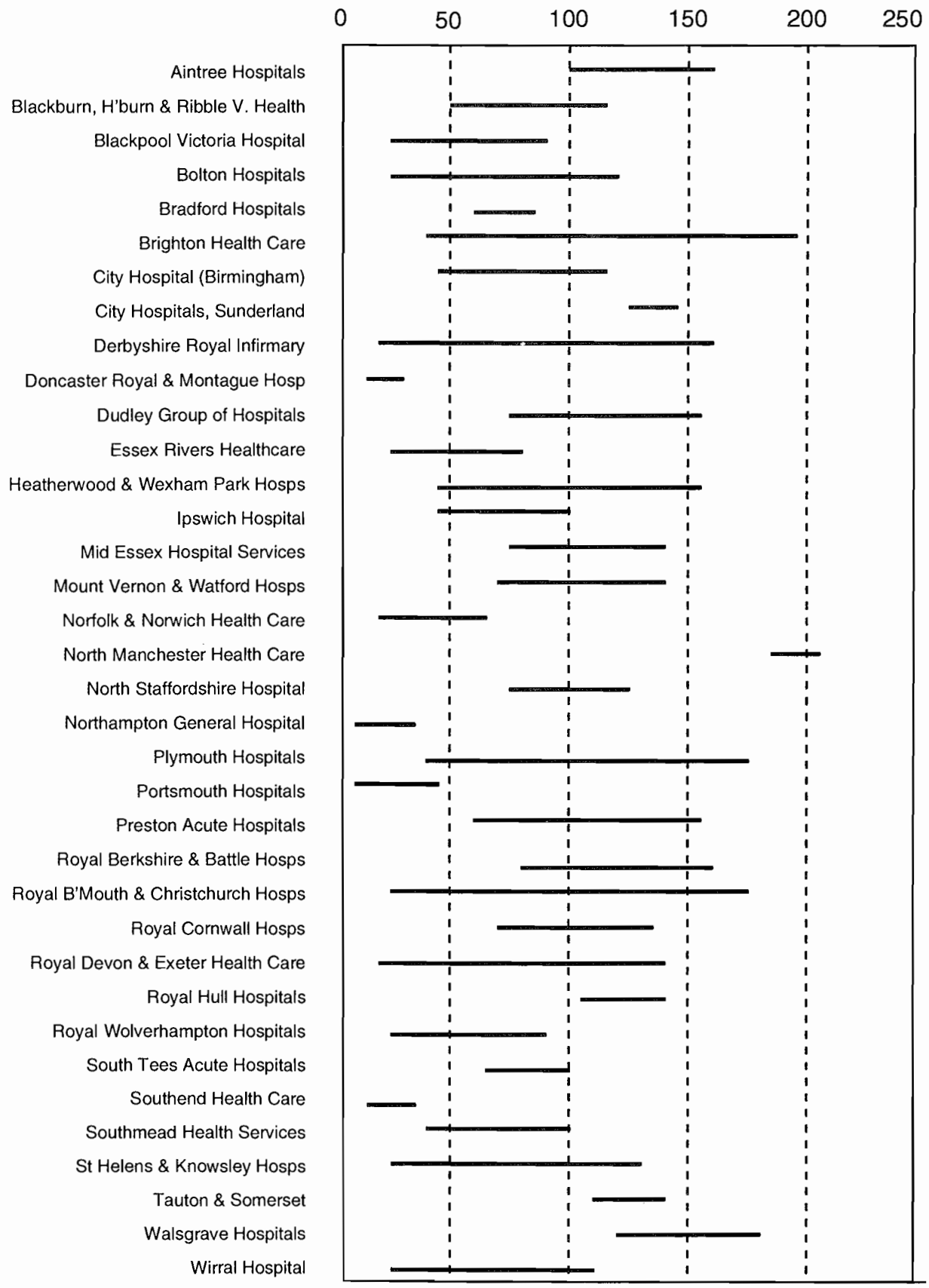

Figure A4 Acute London

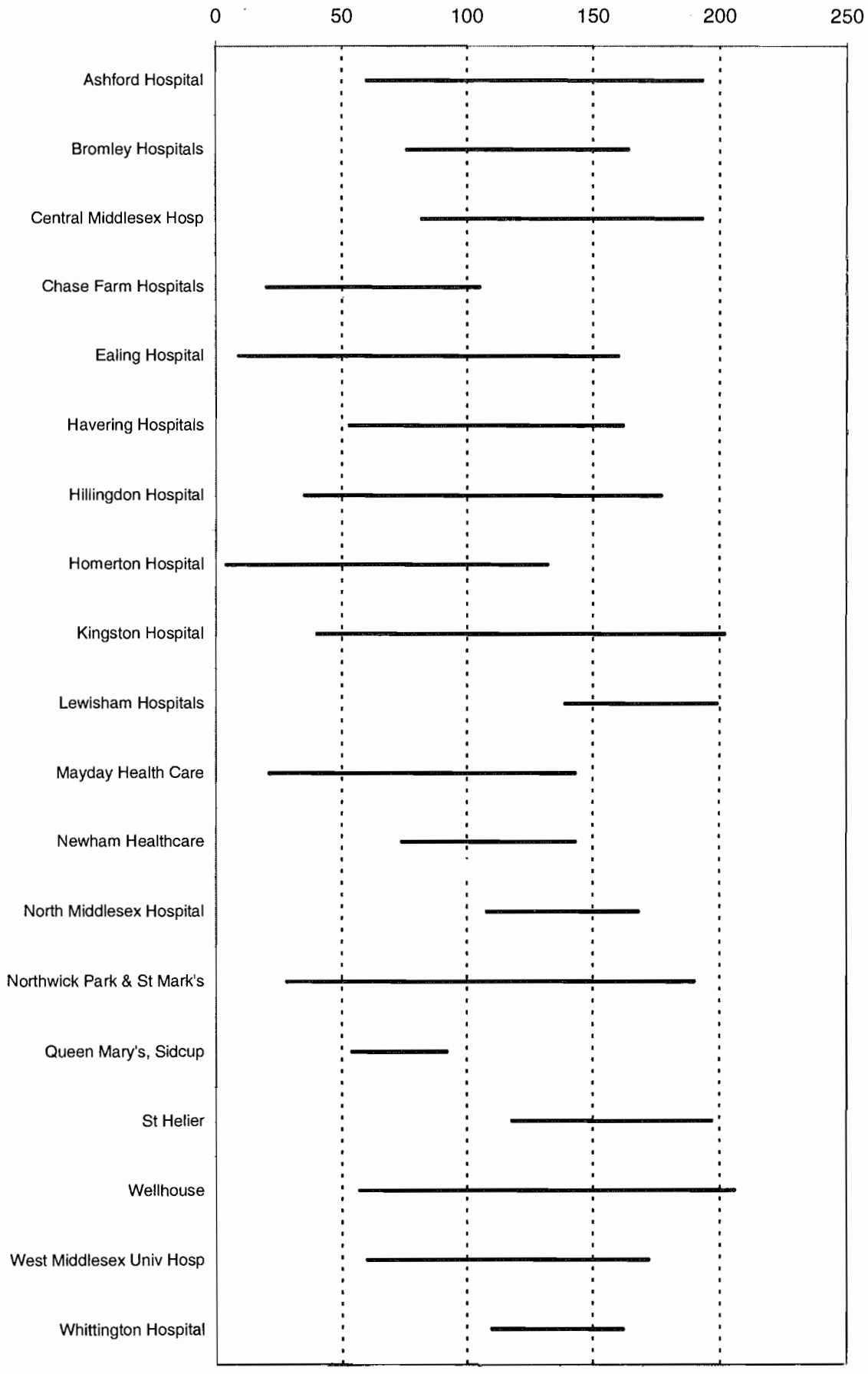

Figure A5 Specialist

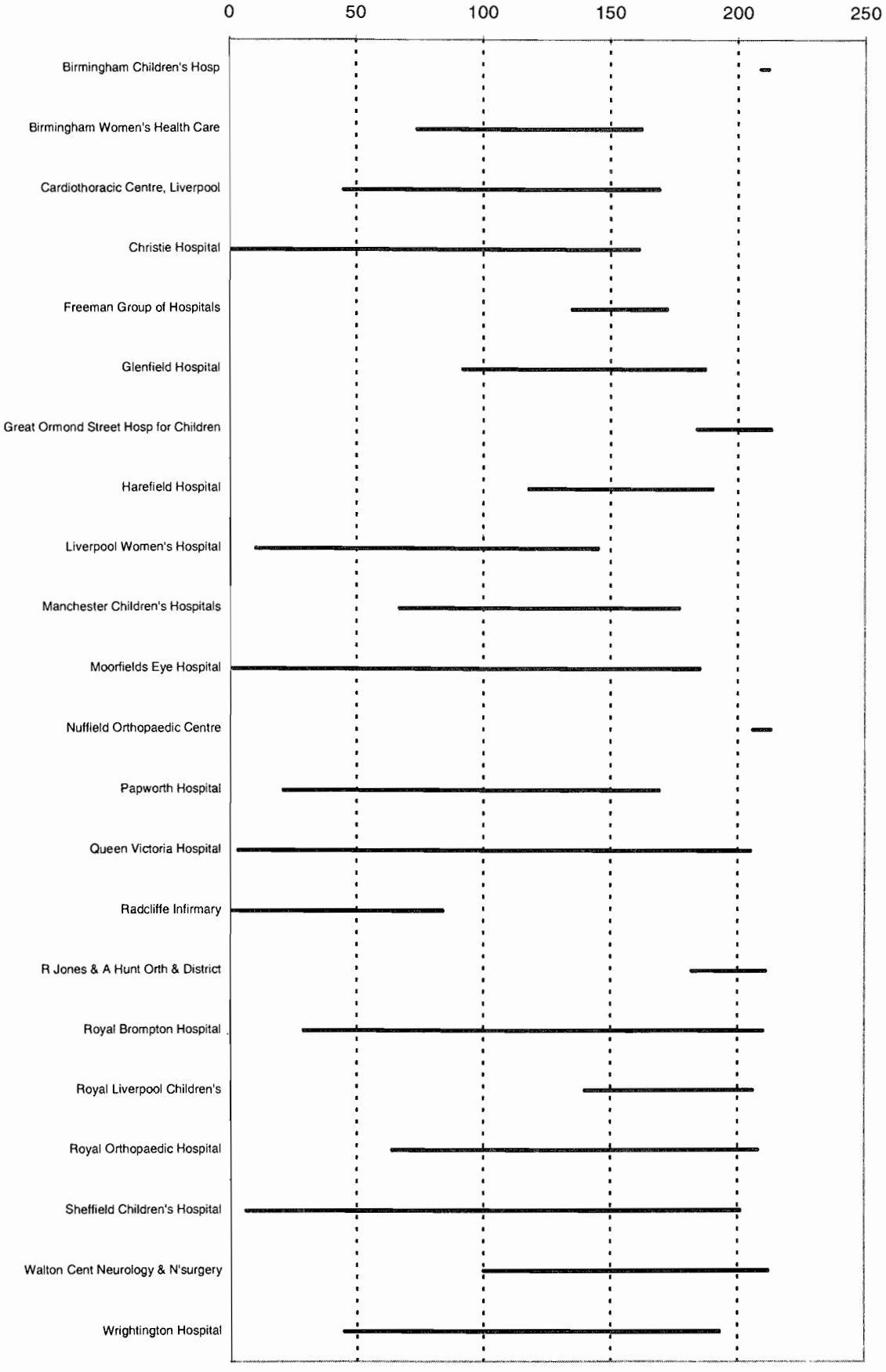

Figure A6 Acute Teaching

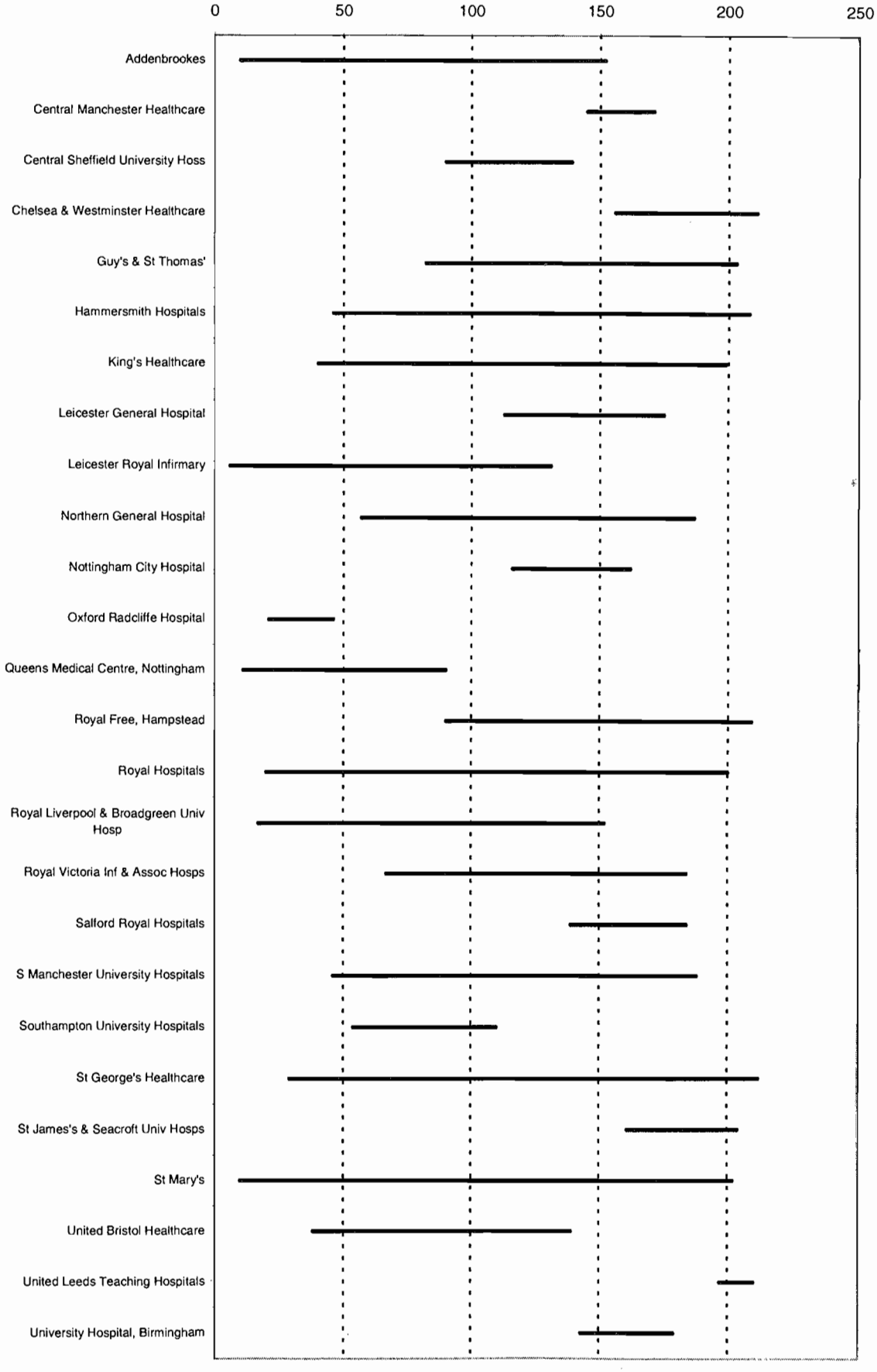

Figure A7 Small/Medium/Large Multi-service

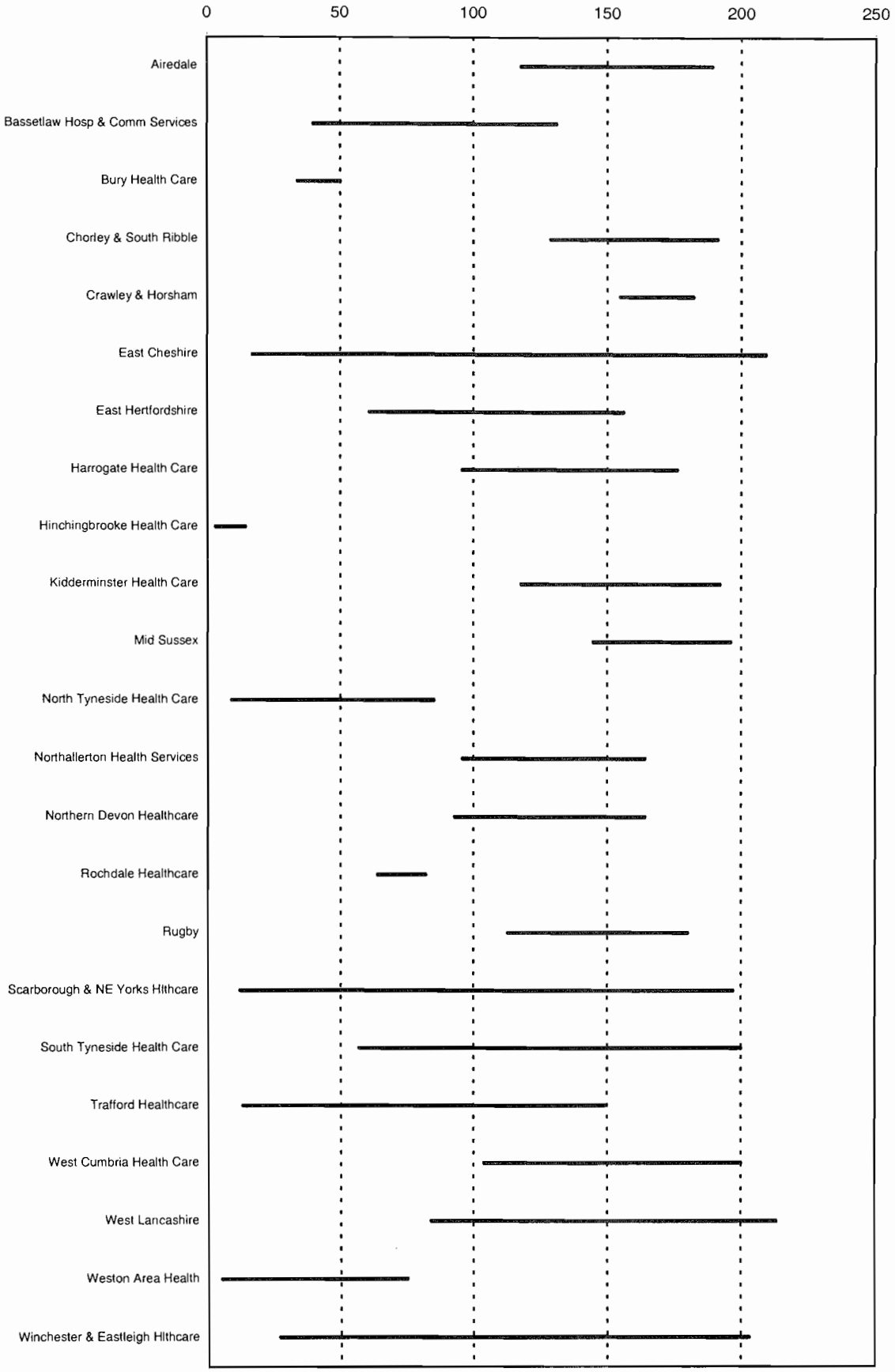

Figure A8 Very Large Multi-service

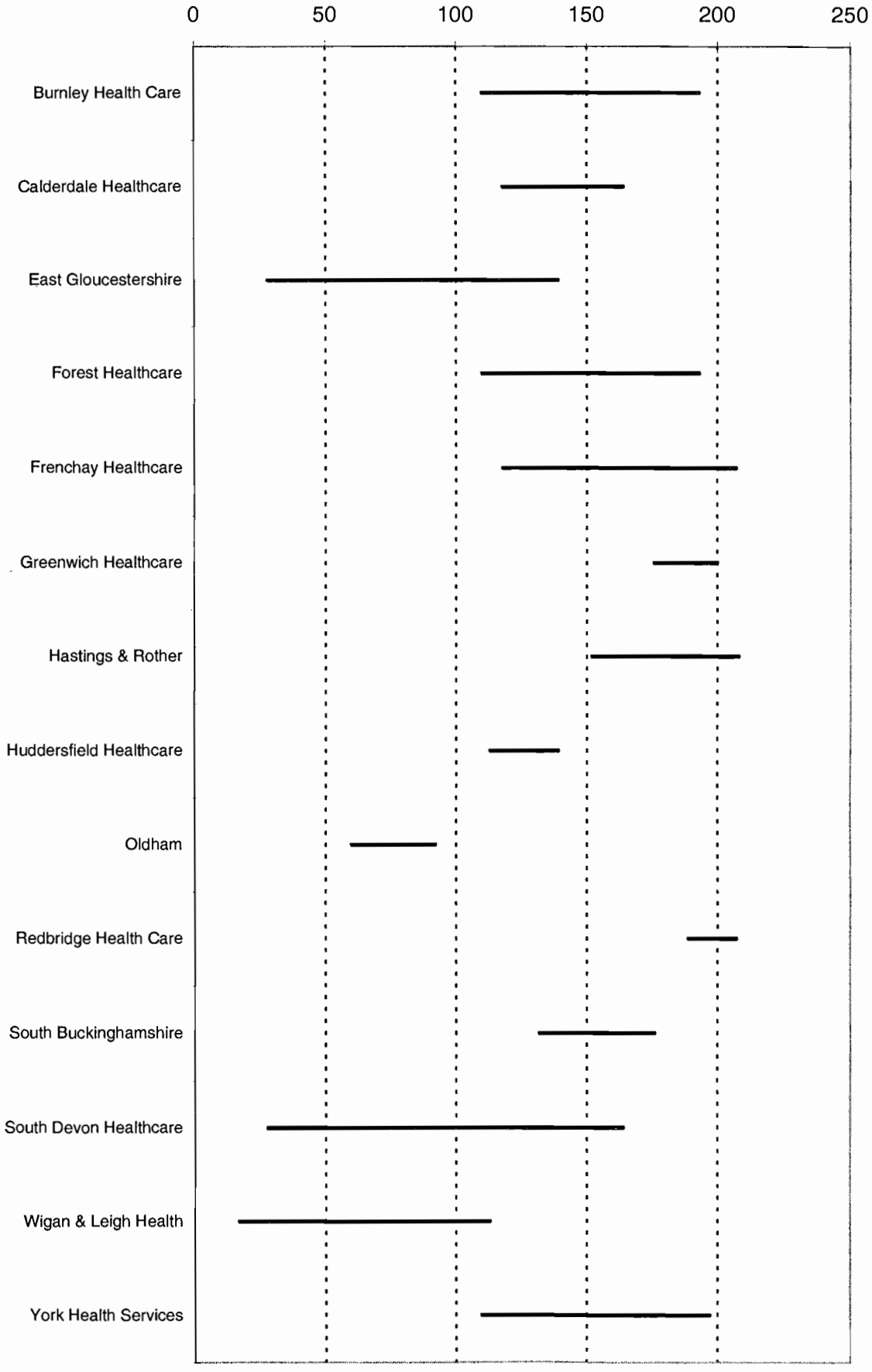

Figure B1 Small/Medium Acute Outside London

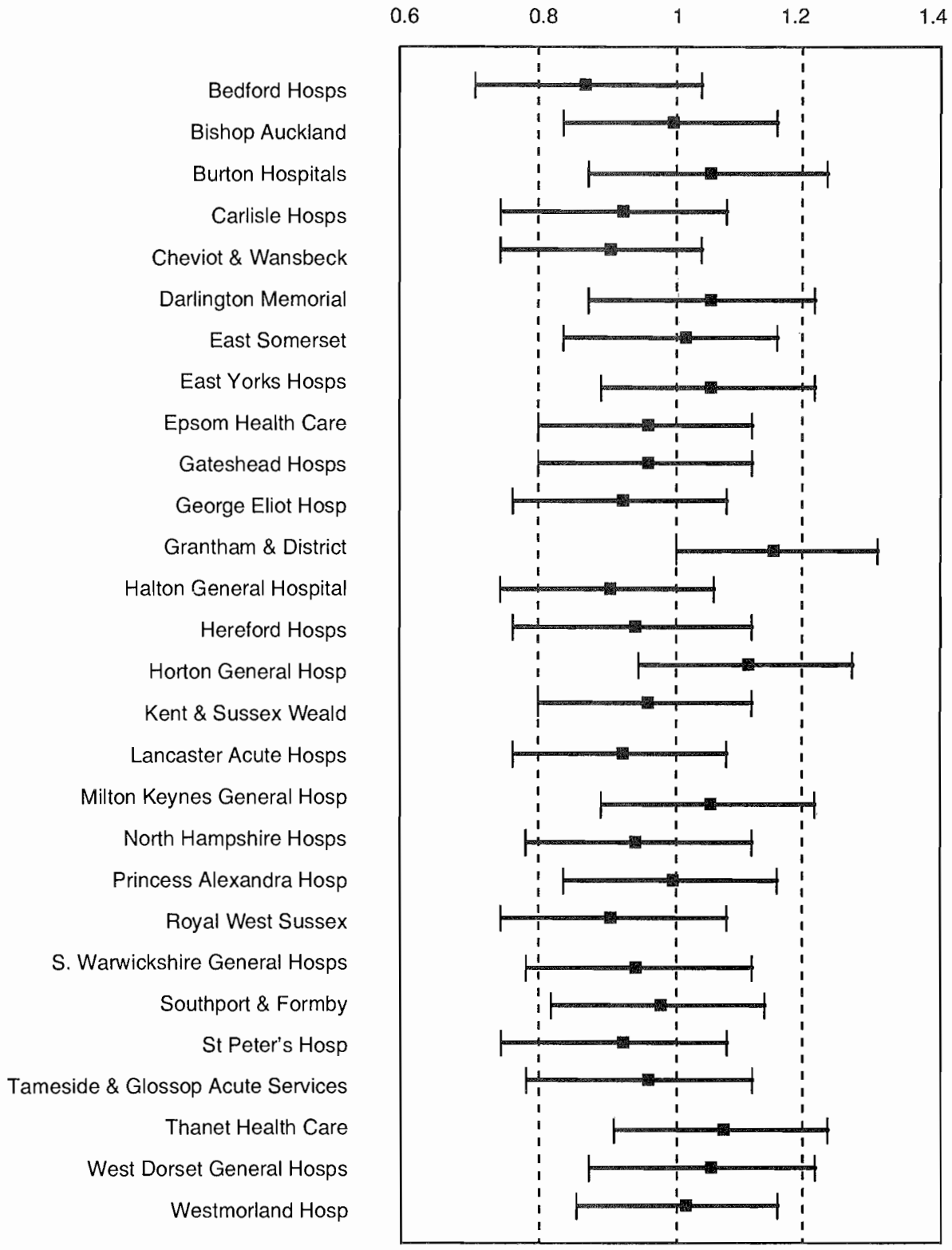

Figure B2

Large Acute Outside London

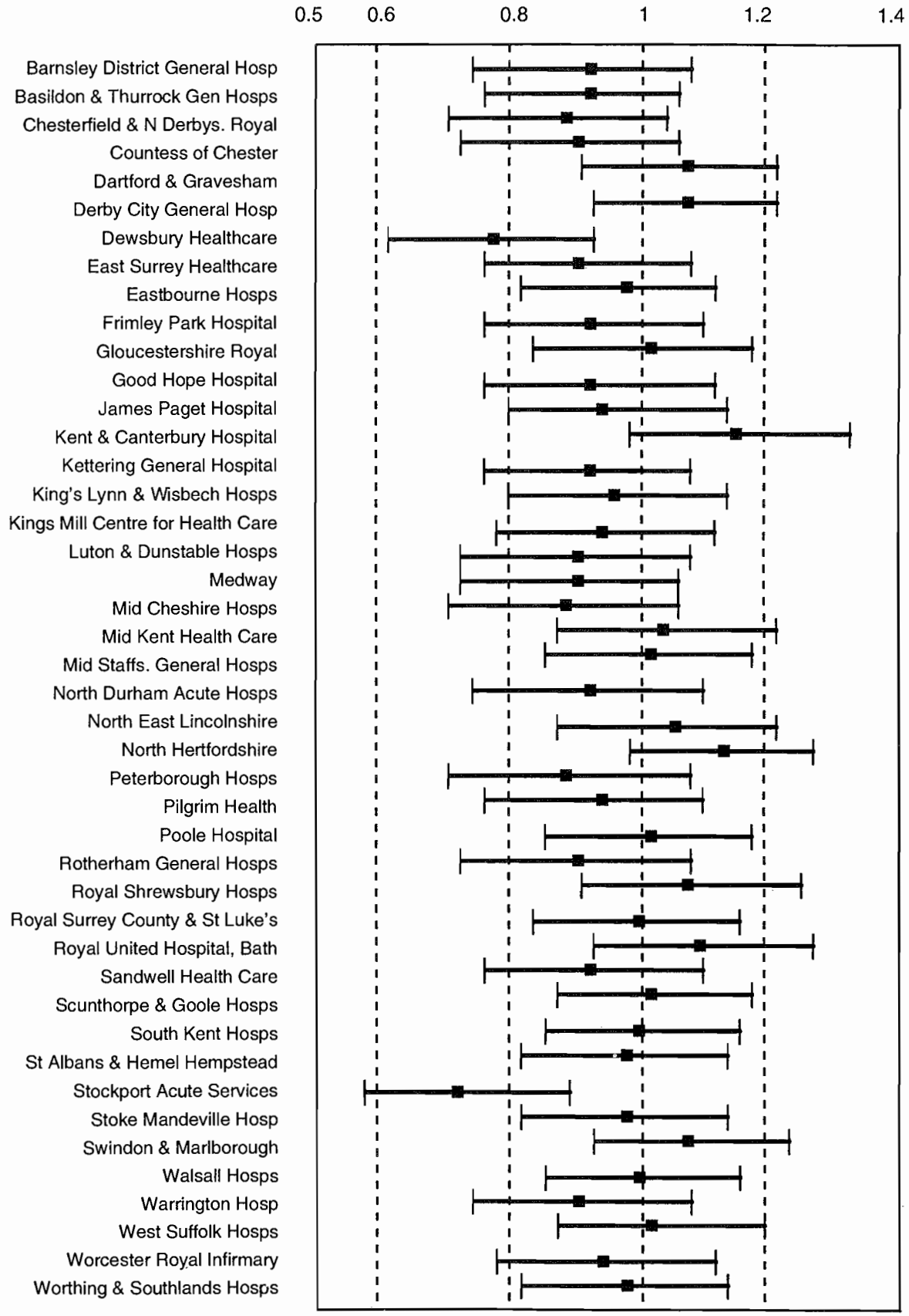

Figure B3 Very Large Acute Outside London

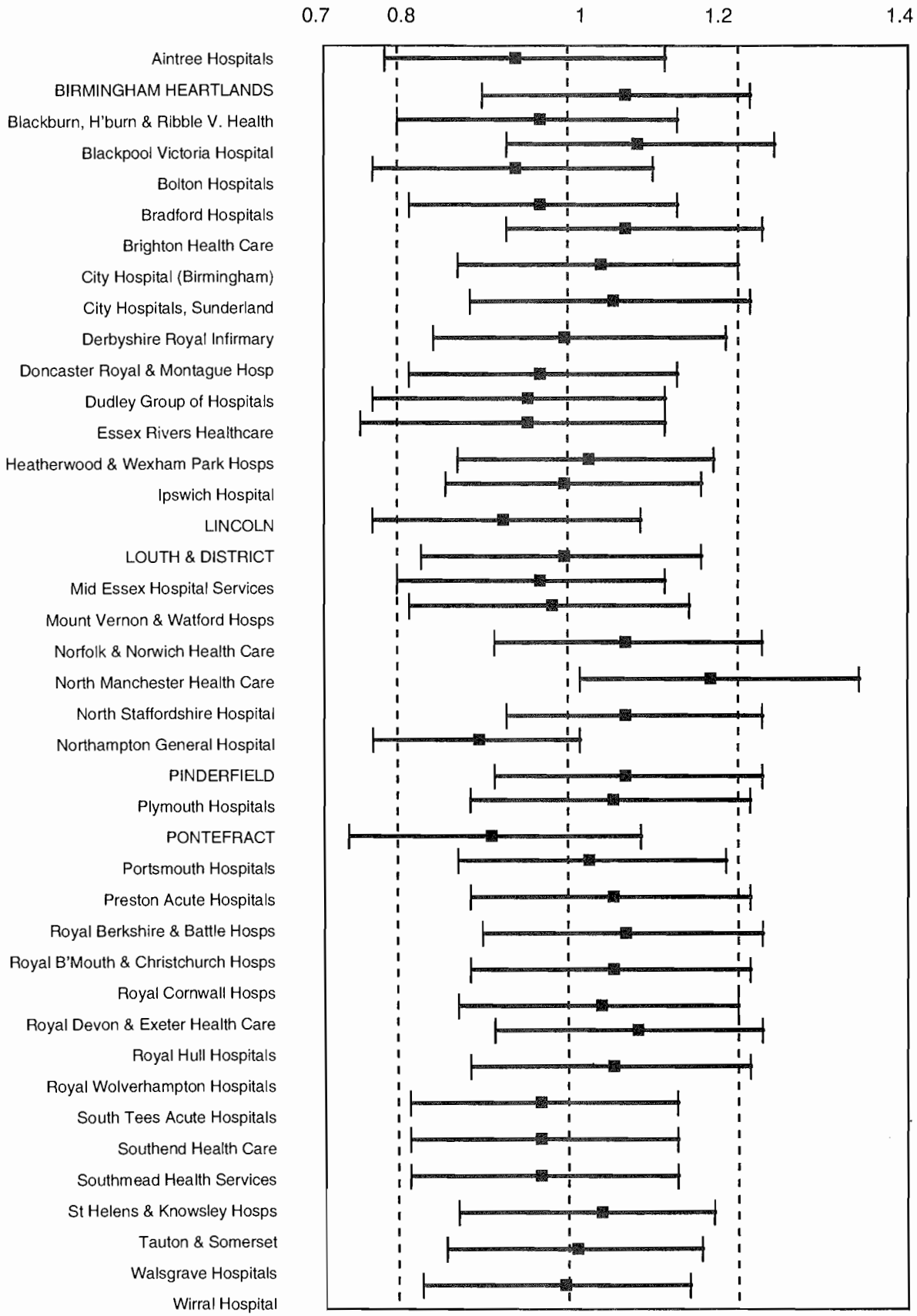

Figure B4 Acute London

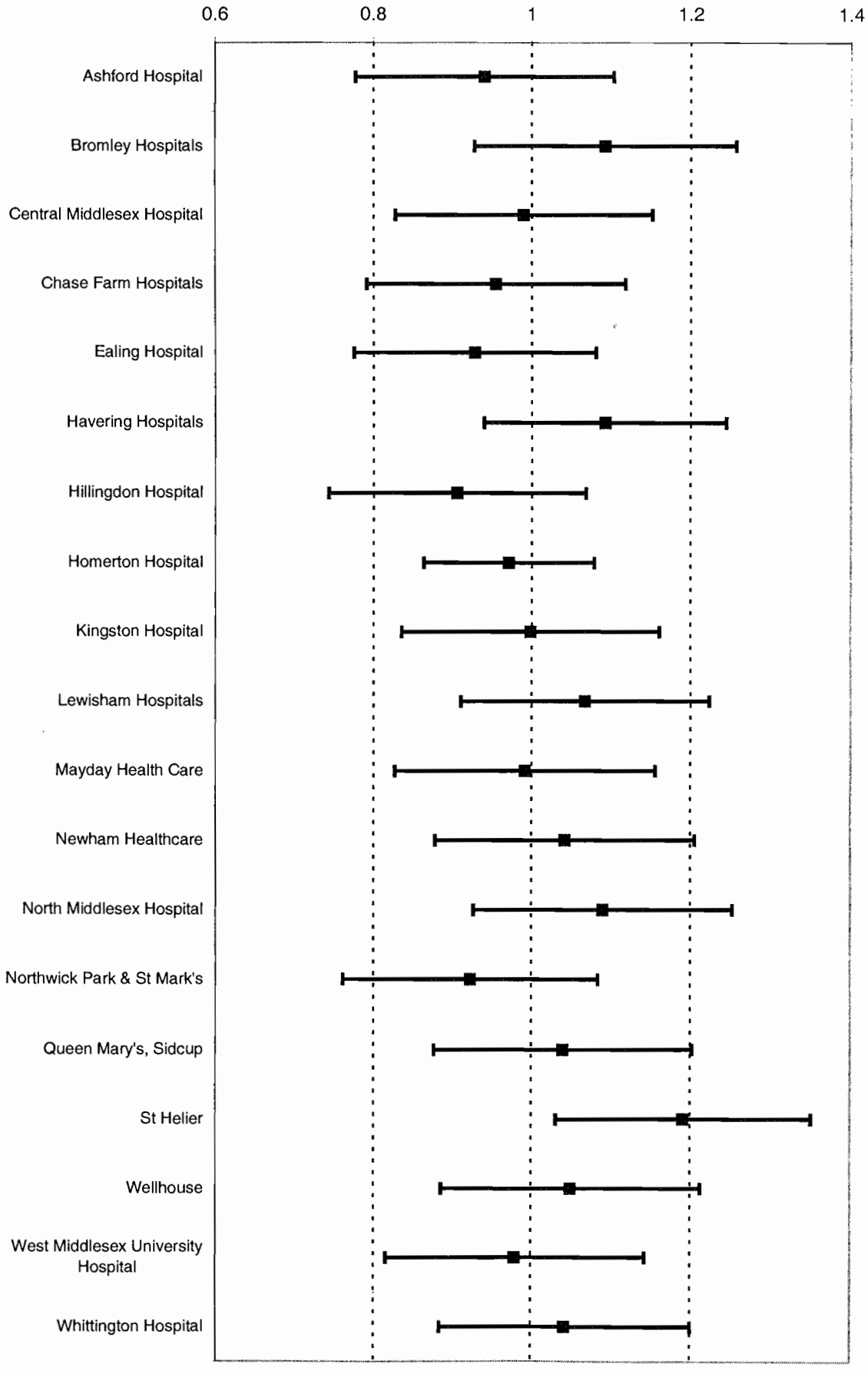

Figure B5 Specialist

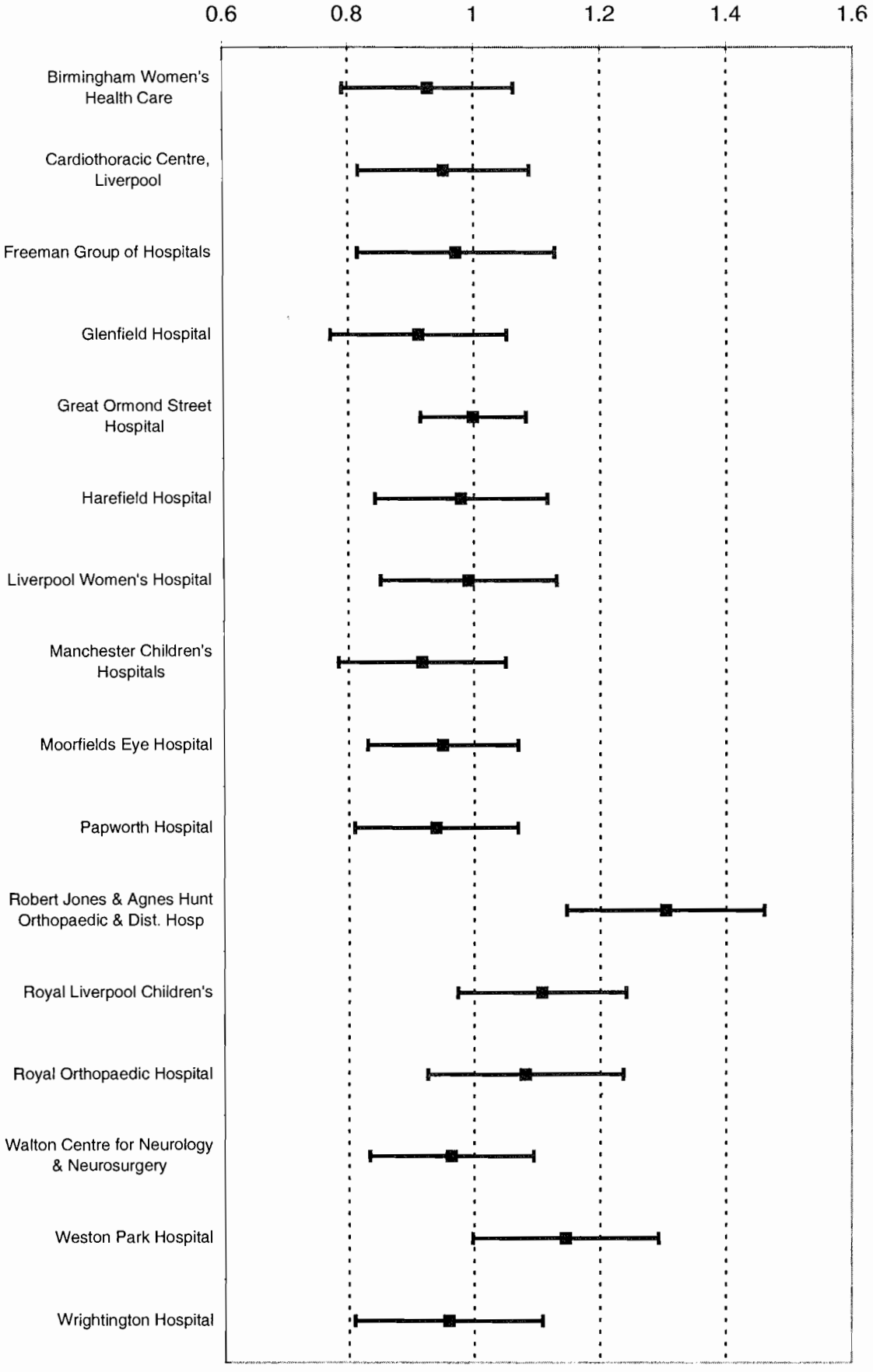

Figure B6 Acute Teaching

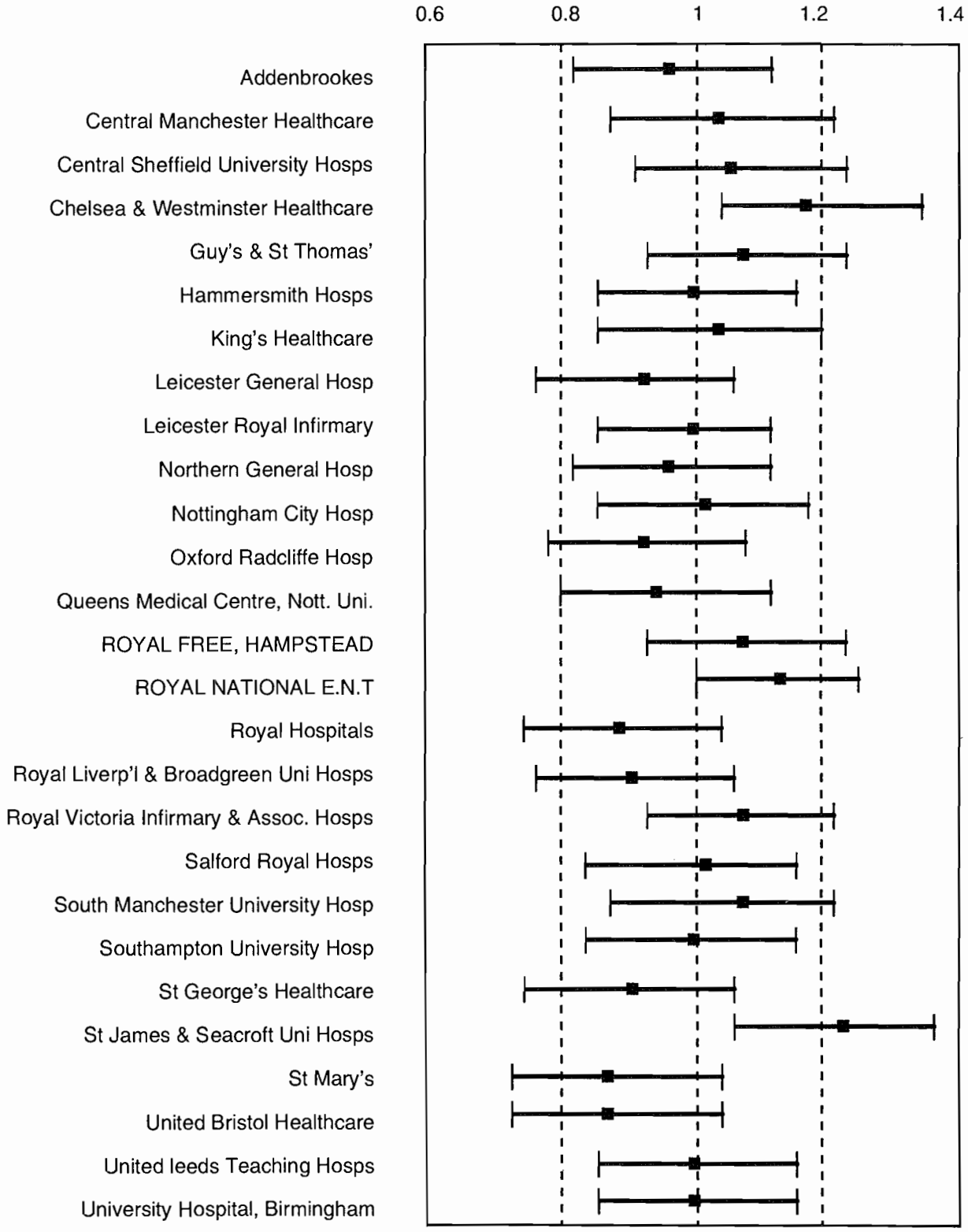

Figure B7 Small/Medium/Large Multi-Service

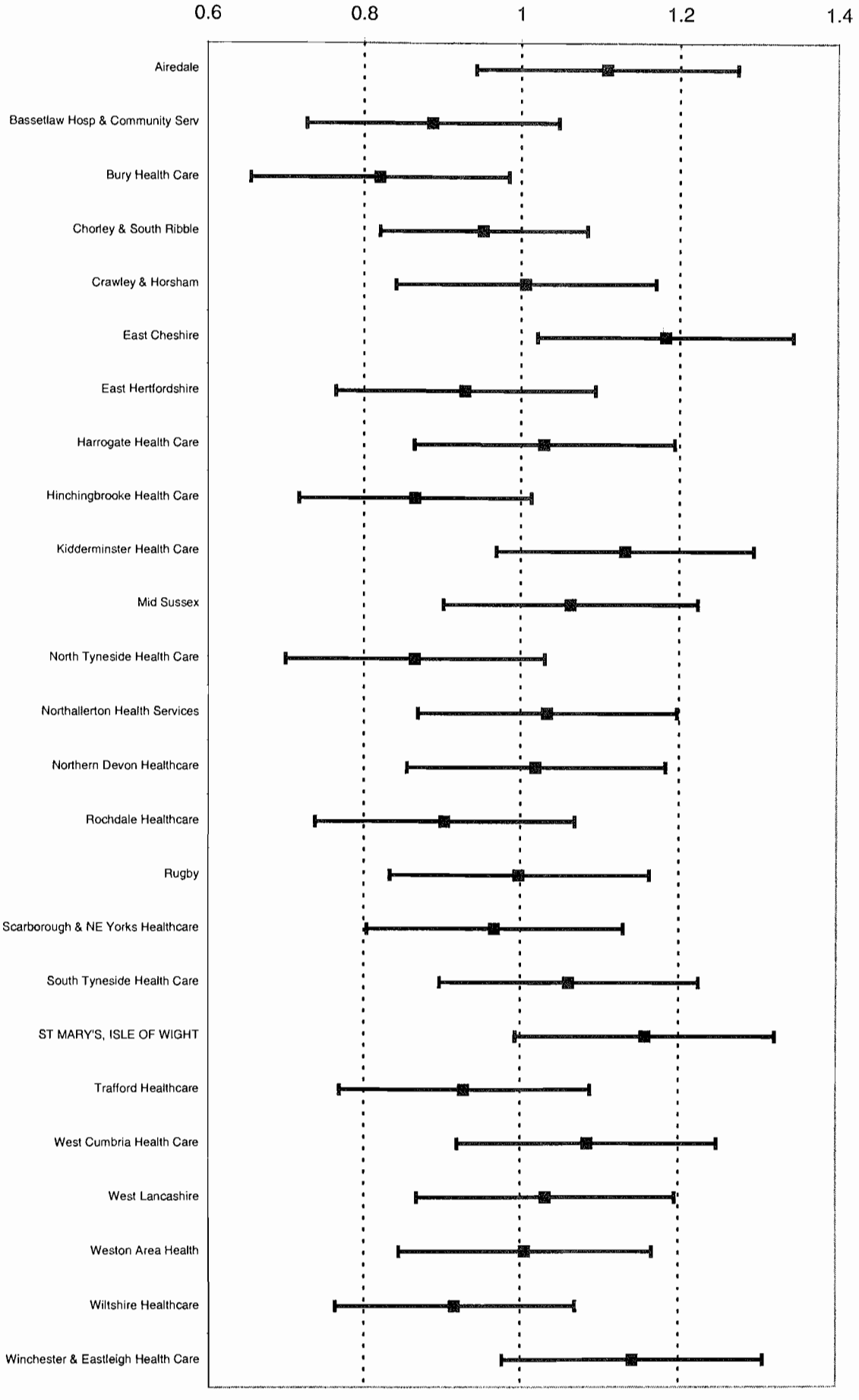

Figure B8 Very Large Multi-Service