THE UNIVERSITY of York

CENTRE FOR HEALTH ECONOMICS

Towards Panel Data Specifications of Efficiency Measures for English Acute Hospitals

Neil Söderlund
Rowena Jacobs

DISCUSSION PAPER 185

Towards Panel data Specifications of Efficiency Measures for English Acute Hospitals

Neil Söderlund

Rowena Jacobs

June 2001

TABLE OF CONTENTS

ABSTRACT

This paper reports work undertaken for the Department of Health to explore different approaches of measuring hospital efficiency. The emphasis throughout is on developing adjusted cost-efficiency measures in line with NHS Trusts performance objectives. Previous work described the derivation of three residual-based cost indices (CCI, 2CCI and 3CCI), each with increasing adjustment in terms of case mix, factor prices and environmental factors for a single year's data (1995/6) (Söderlund & van der Merwe, 1999). This study explores further options based on the previous work by: (1) supplementing hospital level with specialty level data; (2) studying a 4-year panel from 1994/5 to 1997/8; (3) estimating models with non-symmetric error terms and including Trust-specific effects when measuring inefficiency.

Although the paper argues that panel data models may have certain advantages over cross-sectional ones, the results suggest that data pooling across years provide robust parameter estimates. Longitudinal fixed effect models may however be useful to construct efficiency indices while stochastic frontier models have the advantage of taking account of random noise. Specialty level models proved inferior to whole hospital estimations.

The paper argues that the degree of variation between hospitals in terms of efficiency is not that great and scope for efficiency enhancement is primarily attainable by optimising capacity and activity levels in the long run. Increased activity levels may however have adverse consequences such as increased hospital infection rates, poorer quality of care and a lack of capacity to deal with emergency demand. The paper argues that the Department of Health might consider a shift from the adjusted cost index approach used in this normative benchmarking framework to the more conventional efficiency analysis approach using a total cost function, and more flexible functional forms, allowing for a more defensible interpretation of the residuals as inefficiency.

1. INTRODUCTION

Several initiatives have been undertaken by the English Department of Health to provide feedback to NHS Trusts (hospitals) on their performance, as well as to set targets for the coming years. This paper is part of this performance improvement process and builds on the results from earlier work based on performance assessment of acute Trusts (Söderlund & van der Merwe, 1999). The objective of this paper is to compare the results of different approaches to statistical cost benchmarking of acute hospitals and to explore various options for routine analysis of performance measurement. In particular the paper examines whether there are any benefits in using panel data techniques as opposed to cross-sectional analyses in obtaining efficiency estimates.

In most of the cost benchmarking approaches taken by the Department of Health, the common theme is to develop an index that would be comprehensive, transparent, accurate and credible to health service managers. The index represents some form of actual over expected (national average) cost for a given amount of activity. The indices are then statistically adjusted to account for various cost drivers over which hospitals have little control. The unexplained variation in the models is then deemed to represent inefficiency, with cost reduction targets based on the model estimates. In general this approach has several advantages in so far as it may meet some of the criteria of being comprehensive and understandable to managers. However, there are also several theoretical concerns with this approach, including the interpretation of the residuals as inefficiency when the model does not really represent a traditional cost function.

Within this framework, however, efficiency analysis of decision-making units can be used for a variety of purposes (Guiffrida & Gravelle, 1998). Firstly, it can be a means of monitoring the performance of a large number of units, each producing multiple outputs. This may serve to highlight units worthy of further investigation, rather than providing definitive data on performance. For instance, for monitoring or filtering purposes it may not be necessary to agonise over the finer details of statistical methods. Secondly, efficiency analysis may adopt the more definitive stance where differential efficiency estimates are used to produce league tables of performance which are then used in a target-setting process. Thirdly, efficiency analysis can generate information on the cost or production frontier to assist in planning new developments or methods of service delivery, for example to determine the most efficient scale of production. Finally, these analyses may be used to assess the controllable determinants of efficiency, and thus in the long run, lead to policies and practices that improve efficiency. This study has assumed, by and large, that the first objective underlies NHS intentions to compare acute hospitals, although the latter objectives are also touched on in the paper.

The Comprehensive Spending Review (CSR) in 1999 confirmed the government's commitment to improving productivity in all public services through Public Service Agreements (PSAs). As part of the CSR, the NHS was set a target of efficiency gains at 3 percent of Health Authority unified allocations in each of the three CSR years. Through the PSA, efficiency savings in the NHS were largely to come from reductions in Trust unit costs through cost improvements and increased activity

(Noble & Lawrence, 2000). As a result, various initiatives have focused on setting Trust unit cost targets. In 1998 the first National Schedule of Reference Costs and the Reference Cost Index (RCI) were published comparing casemix adjusted Trust costs (and, therefore, implicitly Trust efficiency). The RCI, however, only covered surgical inpatient procedures and accounted for around 20 percent (by cost) of Trust activity. This was not a sufficiently broad base from which to make meaningful comparisons or to set targets. Hence routine Trust returns were used to extend coverage to inpatients, outpatients, day cases and Accident and Emergency (A&E) in all general and acute and maternity specialties, resulting in the RCI+ index (Reference Cost Index +) used to derive Trust targets for 1999/2000.

At the same time, three unit cost indices were developed (CCI, 2CCI and 3CCI) (described more fully in Söderlund & van der Merwe, 1999). These were based on a single cross-section with each cost index making increasing allowance for contextual variables. The indices were more complex, but were felt to be fairer and more accurate measures than the RCI+. In particular the new cost indices were based on complete spells of care rather than individual episodes (Finished Consultant Episodes), constructed from routine data sources and took into account, in addition to casemix, several additional explanatory factors likely to affect efficiency such as capacity, configuration, scale and scope of activity.

However in setting targets for 2000/01, the Department of Health felt that the unit cost measure should be less complex, while still maintaining many of the features of the CCIs. The new index, TUC2000 (Trust Unit Cost 2000) was a hybrid of the RCI+ for 1998/99 but incorporated more RCI data (for which coverage has now been extended to General Medicine). It is effectively a ratio of actual over expected RCI+ and has been used for the most recent target setting process.

This paper builds on earlier work describing the derivation of the CCIs (Söderlund & van der Merwe, 1999). The aim is to explore whether the extension from a cross-section to a panel offers any benefits in the security of parameter (and efficiency) estimates. The feasibility of using more sophisticated statistical adjustments for case mix and other cost drivers is also explored. This paper extends the earlier work by:

- Enlarging the data set to a 4-year panel from 1994/5 to 1997/8.
- Pooling data from multiple years and from multiple specialties to achieve more robust estimates of a provider's performance in the medium term.
- Testing the appropriateness of additional estimation methods including frontier-type models.
- Estimating longitudinal models on all 4 years of data.

This paper is divided into 3 sections. The methods section describes the data, modelling options, the various regressions and the derivation of the cost indices as well as some estimation issues. The results section describes the regression results from the various models and the efficiency scores. Finally, the paper concludes with a discussion of each method, the results obtained and future options for hospital benchmarking.

2. METHODS

2.1 The data

The analysis is based primarily on data from the Hospital Episode Statistics (HES), and hospital accounting returns (TFR1, 2 and 3, and TAC). In addition, particular variables were drawn from one-off surveys of NHS Trusts and data compiled for previous research studies. Data from the various sources were combined into a standard format and Trusts with incomplete data excluded from further analysis[1].

The basic unit of analysis used in this paper is the Trust. Although one may wish to produce efficiency scores at the level of individual specialties within Trusts (and the Department of Health have indeed sought to do this), their validity and usefulness are questionable for a number of reasons:

- Data are not ideal for estimating specialty level models. Many of the variables, such as the capacity, teaching, research, competition and specialisation terms were only available at the whole hospital level, although they clearly would be more meaningful at the specialty level for such models. In addition, some data collected at the specialty level, especially expenditure data, might be allocated arbitrarily or inconsistently between specialties, especially where shared inputs are concerned.
- There are doubts about specialties being valid independent units for analysis since it is unclear what the economic incentives might be that underlie behaviour at the level of specialties. Some terms, such as variable costs and activity, may have meaning at the specialty level. Others, such as hospital size, may not. One could assume that inefficiency at the hospital level due to these factors is ‘passed down’ to the specialty level, but its proportional allocation would be arbitrary at best. It is possible that specialty level disaggregation allows for better case mix adjustment. Other research suggests, however, that the ability of specialty to predict resource use for acute care is poor (Söderlund *et al*, 1996). Since specialties in essence compete within Trusts for resources, one would in principle have to model a budget allocation process between specialties. This would require a series of equations optimising specialty budgets within a global Trust budget constraint, which would, given the above reasons, prove very difficult.
-

Therefore, specialty-level data was aggregated to Trust level and regressions for individual specialties were not conducted. We did however conduct hospital level analyses using specialty level data, where specialty was used as a regressor, rather than estimating separate models for each specialty. The aggregation of specialties is described in Appendix 1.

The derivation of each of the variables used in the modelling is described in Appendix 2. Some data limitations should be borne in mind. Inconsistencies in data may arise from different ways in which Trusts classify inpatients, outpatients and day cases and these may in turn influence the resulting efficiency scores. However this may be more problematic in assessing performance within specialties than at Trust level, and the extension of Healthcare Resource Groups (HRGs) to outpatients and A&E in the future would reduce this classification problem. Trusts also may have different ways in which they allocate costs across specialties, for example the cost of consultants

[1] Approximately 99 hospital-years’ worth of data was excluded, leaving a sample of 951 years of data.

who work across specialties. Again this would affect the reliability of specialty estimates more than Trust level estimates. Finally, data cannot yet make allowance for cost shifting from Trusts to community agencies, patients, families and social services. For example early discharge may help to lower a Trust's unit costs and improve its efficiency score by shifting the care burden into the community. Indicators such as readmission rates may help to monitor this in future.

2.2 Derivation of the cost indices

Previously three separate cost indices have been developed for the Department of Health to produce efficiency rankings for Trusts in order to benchmark their performance based on their productivity scores (Söderlund & van der Merwe, 1999). This analysis uses the same variables and statistical adjustments to the deterministic cost index, as briefly described below.

The CCI is a deterministic cost index of actual divided by expected costs, where expected costs are average national costs for each type of activity and include case-mix adjusted inpatient, first outpatient and A&E attendances. 2CCI and 3CCI are long run (partially adjusted) and short run (fully adjusted) indices regressed against the CCI with increasing numbers of explanatory variables. 2CCI takes factors into account such as additional adjustments for case mix, age and gender mix, transfers in and out of the hospital, inter-specialty transfers, local labour and capital prices, competition and teaching and research outputs for which Trusts might be over or under compensated. The 3CCI makes additional adjustments, over and above those in the 2CCI, for hospital capacity, including number of beds, and number of sites, scale of inpatient and non-inpatient activity and scope of activity. Therefore this index tries to capture institutional characteristics amenable to change in the long, but not the short run.

The use of an adjusted cost index[2] as the dependent variable is rather constraining from an econometric perspective as it fixes certain adjustments *a priori*, prevents the proper investigation of scale and scope economies and interaction effects, and makes the interpretation of the coefficients on many regressors difficult. Nevertheless, it was felt that such an approach would be more intuitively appealing to NHS managers than a less constrained regression on total cost. However, this does complicate the interpretation of the residuals as inefficiency measures, since the CCI is essentially an index standardised around one and not a cost function where the residuals may represent deviations from the frontier. Therefore the analysis may be best interpreted as explaining variations in the CCI which the Department of Health deems to be broadly similar to the RCI+.

The short run (3CCI) and long run (2CCI) adjusted indices are calculated from the model as follows (the preferred log-log form of the model is shown, although they can just as easily be calculated from the linear additive form):

3CCI: $$I_{it}^{SR} = \exp(\ln CCI_{it} - \beta \ln X_{it}' - \gamma \ln Z_{it}' - \mu T_t - \alpha) \qquad (1)$$

[2] It should be noted however that because of the log-log formulation chosen, the model essentially amounts to a total cost function since $\ln(CCI i) = \ln(C i) - \ln([(\underline{IC}*IP i*HI i)/(\underline{HI}*\underline{IP})] + \Sigma_j OP ij*\underline{OC j}/\underline{OP j} + AE i*\underline{AC}/\underline{AE})$.

2CCI: $I_{it}^{LR} = \exp(\ln CCI_{it} - \beta \ln X_{it}' - \mu T_t - \alpha)$ (2)

where:

ln indicates natural logarithm, and exp indicates the corresponding antilog

I_i^{SR}	=	Normally distributed short run efficiency indicator for hospital *i* averaged over all time periods *t*
I_i^{LR}	=	Normally distributed long run efficiency indicator for hospital *i* averaged over all time periods *t*
CCI_{it}	=	Cost index for hospital *i* in period *t*
X_{it}	=	Vector of "production theoretic" regressors (outputs and factor prices)
Z_{it}	=	Vector of institutional characteristics amenable to change in the long, but not the short run
T_t	=	*t*-1 period specific dummy variables
$\beta, \gamma, \mu, \alpha$	=	Parameters to be estimated

The variables in these benchmarking regressions are shown in Table 1 and described more fully in Appendix 2.

Many of the independent variables used in the model are divided through by scale or capacity factors (for example episodes per spell, FCEs per spell, students per spell). The deflation provides a more precise estimate of the individual effects of the regressors on the dependent variable and also reduces heteroskedasticity (where OLS would otherwise place more weight on observations with large error variances).

It was decided that no specifications with higher order terms would be considered. This is contrary to most of the recent theoretical literature on hospitals (and multiproduct industries in general) (Berry, 1970; Braeutigam & Daughety, 1983; Breyer, 1987; Butler, 1995; Friedlander & Spady, 1981; Hornbrook & Monheit, 1985). It also allows only fairly simple treatment of scale and scope issues, and is not strictly derived from any underlying production function. The reasons for the simplification made here were pragmatic rather than theoretical. Firstly, the coefficients on models with only first order terms are far easier to interpret than those of more flexible functional forms. Furthermore, too flexible a functional form might be counterproductive in this exercise given its normative set of objectives and the fact that residuals are interpreted as efficiency scores. Often when more flexible functional forms such as translog models are used, they turn out to be near-deterministic and there is no residual left with which to study inefficiency. Since the primary objective of this efficiency analysis was not to generate information on the cost frontier for determining scale or scope efficiencies, but for monitoring and filtering purposes, this approach seemed justified. Finally, specification tests on the models with only lower order terms generally indicated that they were not mis-specified.

Three functional forms were considered sufficiently simple and easily interpretable for consideration for the basic model (with all data years pooled): the linear additive, semi-log and log-log functional models. We had no theoretical reason to favour one above the others. While parameter estimates and model fit were similar for all three, the log-log form showed consistently better results as judged by the RESET test (Ramsey, 1969) and adjusted R-squared and this form was used. The results for each are shown below.

TABLE 1: Explanatory Variables

Variable name	Log-transformed	Description
Dependent variable		
COSTINDX	LNCSTNDX	CCI cost index
Long run adjusters (2CCI and 3CCI)		
INTERCEP		Intercept
EP_SPELL	LNEP_SPL	Episodes per spell
TRANSIPP	LNTRANSI	Transfers in to hospital per spell
TRANSOPP	LNTRANSO	Transfers out of hospital per spell
EMERGPP	LNEMERGP	Emergency admissions per spell
FCEINPP	LNFCEINP	Finished consultant episode inter-specialty transfers in and out of specialty
OPNPP	LNOPNPP	Non-primary outpatient attendances per inpatient spell
EMERINDX	LNEMERIN	Standardised index of unexpected emergency admissions/total emergencies
HRGWTNHS	LNHRGWT	HRG weight, case mix index
PROP15U	LNPROP15	Proportion of patients under 15 years of age
PROP60P	LNPROP60	Proportion of patients 60 years or older
PROPFEM	LNPROPFM	Proportion of female patients
STUDENPP	LNSTUDPP	Medical student whole time teaching equivalents per inpatient spell
RESEARPC	LNRESEAR	Percentage of total revenue spent on research (estimated 1995)
MFF_COMB	LNMFFCOM	Market forces factor - weighted average of staff, land, buildings and London weighting factors
HERF15	LNHERF15	15 mile Herfindahl (competition) index
POPDENS	LNPOPDEN	Population density
Short run adjusters (3CCI only)		
HESSPNHS	LNHESSP	Total inpatient spells by NHS patients
TOTOP1	LNTOTOP1	Total primary outpatient attendances
A_E1	LNA_E1	Total primary A&E attendances
AVBEDS	LNAVBEDS	Average available beds
HEATBED	LNHEATBD	Heated volume per bed
SITES50B	LNSITES	Sites with more than 50 beds
ITINDX	LNITINDX	Scope / specialisation (information theory) index
MERGED	Not log transformed	Dummy indicating whether a Trust had recently been part of a merger / reorganisation or not
Other variables		
YEAR45		1994/5 dummy variable
YEAR56		1995/6 dummy variable
YEAR67		1996/7 dummy variable
YEAR78		1997/8 dummy variable

TABLE 2: Functional Form

Statistic	Linear	Semi-log	Log-log
Adjusted R-squared	.66	.68	.71
RESET – F	2.2	7.0	0.86
RESET - p	.08	.0001	.45

In addition, the log-log form has been adopted for the most recent Department of Health / Audit Commission release of the Reference Cost Indices.

2.3 Modelling options

Despite attempts to limit the range of approaches used by normative exclusions, theoretical convention and specification tests, there remained an array of different approaches to be considered for hospital cost modelling. The issues that need to be considered and the options are shown in Figure 1.

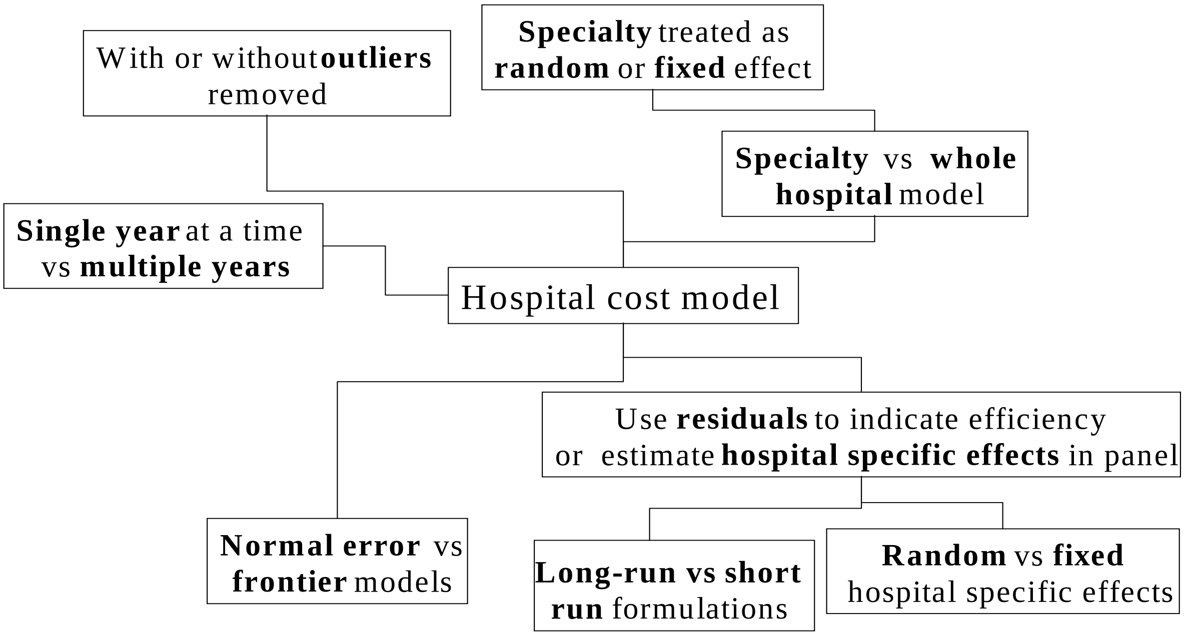


Figure 1: Modelling Option TREE

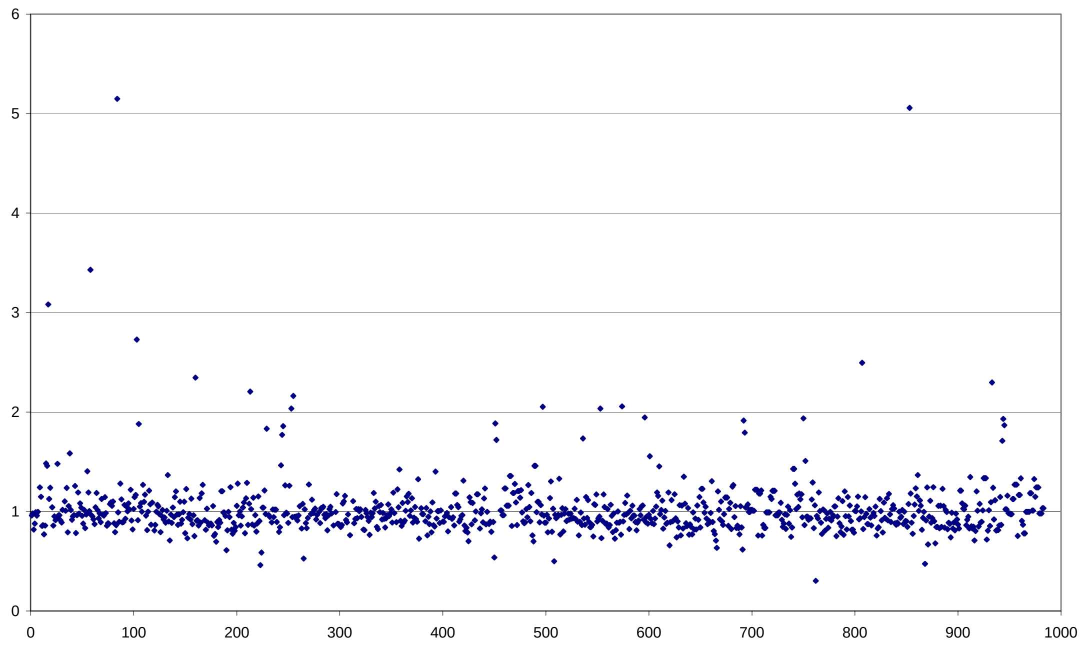
Not every option could be reported. Those that seemed promising are described in this paper. Others were tested but added little to existing models.

Two sets of regressions were run in each case, one on the full sample of Trusts and one excluding the atypical, highly influential, or outlier data points in the model by using the DFITS statistical procedure[3]. By identifying and removing these observations the model was re-estimated for a subset of typical providers, and thus general efficiency norms could be established more accurately without the influence of these atypical Trusts. Therefore, efficiency scores were taken from the outlier-excluded estimate, except for outlier hospitals, where the full sample estimate was used.

Figure 2 shows the distribution of the CCI cost index (with all years pooled) including outliers. It is interesting to note that the outliers as shown here are often specialist hospitals (e.g. neurology, orthopaedics, paediatric hospitals) which suggests insufficient adjustment for specialisation in the CCI. It should be noted however that an (non)outlier as shown on the CCI figure may not necessarily be an (non)outlier by the DFITS procedure as it is based on model residuals.

[3] The DFITS statistical procedure (Belsey, Kuh & Welsch, 1980) was used in these analyses to identify highly influential providers, and exclude these from the estimation of coefficients. A relatively conservative threshold was used to exclude observations: where DFITS > $2 \times (p/n)^{0.5}$ (p = number of parameters estimated; n = sample size).

FIGURE 2: Dispersion OF CCI Around Mean One With Outliers



Alternative estimation techniques were used to test the robustness of rankings generated by a simple OLS model with pooled data. Three basic modifications were tested. The first was the random and fixed effects model (at whole hospital level) which attempts to capture specialty effects by pooling specialty level data. The second modification was the use of non-symmetric statistical techniques, also known as a 'frontier' model. Stochastic frontier approaches allow for both random error, and an asymmetric error term which is assumed to reflect inefficiency, so that Trusts are compared to a statistically determined, rather than a deterministic frontier. The third modification was the extension of the OLS model into random and fixed effects panel models (at whole hospital level) to capture provider effects.

All of the above alternative approaches were applied to the data from which outliers had already been removed.

The following models are therefore presented in this paper:

Model 1: The basic starting model – an **OLS estimation at the whole hospital level**, in both long and short run formulations (equations (1) and (2)), on data with all years pooled and outliers excluded, using the average residual for a Trust over all four years as its efficiency score. The short-run (fully specified model – (1)) form of this was then compared to the following variants, all estimated on outlier-excluded data.

Model 2: Pooled specialty level model where **specialty is treated as a fixed effect**, and efficiency scores are taken as the average model residual.

$$\ln CCI_{ijt} = \alpha + \beta \ln X_{ijt}' + \gamma \ln Z_{ijt}' + \mu T_t' + \phi S_j' + I_{ijt}^{SR} \qquad (3)$$

where terms are as for (1), S_j represents j-1 specialty dummies, and I_i^{SR} is averaged over all specialties j and periods t.

Model 3: Pooled specialty level model where **specialty is treated as a random effect**, and efficiency scores are taken as the average model residual.

$$\ln CCI_{itj} = \alpha + \beta \ln X_{itj}' + \gamma \ln Z_{itj}' + \mu T_t' + \theta_j + I_{itj}^{SR} \qquad (4)$$

where θ_j represents j specialty specific random effects and I_i^{SR} is averaged over all specialties j and periods t.

Model 4: As for Model 1 except a **stochastic frontier**, assuming half-normal error distribution and the average residual taken.

$$\ln CCI_{it} = \alpha + \beta \ln X_{it}' + \gamma \ln Z_{it}' + \mu T_t' + v_{it} + H_{it}^{SR} \qquad (5)$$

where v_{it} is a random error term assumed to be independent of H_{it}^{SR}, a random error term distributed as half normal and averaged across all t to give H_i^{SR}.

Model 5: As for Model 1 except **each Trust is treated as a random effect**, and efficiency scores are taken from the random effect estimates.

$$\ln CCI_{it} = \alpha + \beta \ln X_{it}' + \gamma \ln Z_{it}' + \mu T_t' + \delta_i + \varepsilon_{it} \qquad (6)$$

where δ_i is a provider specific random error term, and ε_{it} is a normally distributed random error term which captures relative costliness of provider i across all time periods.

Model 6: As for Model 5 except **each Trust is treated as a fixed effect**, and efficiency scores are taken from the Trust fixed effect estimates.

$$\ln CCI_{it} = \alpha + \beta \ln X_{it}' + \gamma \ln Z_{it}' + \mu T_t' + \omega P_i' + \varepsilon_{it} \qquad (7)$$

where P_i is a provider specific dummy variable, ε_{it} is a normally distributed random error term, and ω captures relative costliness of provider i across all time periods.

Ramsey's RESET test (Ramsey, 1969) for miss-specification and omitted variables was applied. In addition to conventional indicators of model fit, such as the R-squared, the mean squared error (MSE) and the F-statistic, efforts were made to assess inefficiencies due to multicollinearity and heteroskedasticity using the Variance Inflation Factor (VIF) statistic and the Cook-Weisberg and White's test (Freund & Littell, 1991; Cook & Weisberg, 1983; White, 1980) respectively.

TABLE 3: Parameter estimates From Regressions 1 TO 6.

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Variable	**Estimate**	**Prob > |T|**	**Estimate**	**Prob > |T|**	**Estimate**	**Pr > |t|**	**Estimate**	**Pr > |t|**	**Estimate**	**Pr > |t|**	**Estimate**	**Pr > |t|**
INTERCEP	-0.7365	0.0029	-0.7329	0.0001	-0.6475	0.0011	-0.7823	0.0023	-0.4214	0.2560	6.5609	0.0000
LNTRANSI	-0.0023	0.4508	0.0050	0.0001	0.0050	0.0001	-0.0023	0.5089	0.0012	0.6700	0.0022	0.4920
LNTRANSO	-0.0005	0.8684	0.0049	0.0001	0.0048	0.0001	-0.0006	0.8598	0.0026	0.3270	0.0032	0.2930
LNEMERGP	-0.0014	0.9144	0.0193	0.0001	0.0195	0.0001	-0.0024	0.8792	0.0252	0.0420	0.0705	0.0000
LNFCEINP	-0.0063	0.0098	0.0024	0.0425	0.0024	0.0450	-0.0068	0.0347	-0.0031	0.1670	0.0000	0.9990
LNOPNPP	0.1001	0.0001	0.0374	0.0001	0.0376	0.0001	0.1005	0.0000	0.1342	0.0000	0.1332	0.0000
LNEMERIN	-0.0227	0.0036	0.0235	0.0001	0.0229	0.0001	-0.0218	0.0116	-0.0079	0.1850	-0.0046	0.4250
LNEP_SPL	0.3535	0.0001	0.4856	0.0001	0.4856	0.0001	0.3484	0.0000	0.3392	0.0000	0.3131	0.0000
LNHRGWT	0.0719	0.0444	0.1049	0.0001	0.0927	0.0001	0.0654	0.0777	0.0454	0.2700	-0.2412	0.0000
LNPROP15	-0.0035	0.3403	-0.0020	0.2138	-0.0020	0.2198	-0.0036	0.3466	0.0010	0.7820	0.0099	0.0310
LNPROP60	-0.0321	0.0001	-0.0150	0.0001	-0.0151	0.0001	-0.0323	0.0000	-0.0253	0.0000	-0.0284	0.0010
LNPROPFM	-0.0950	0.0206	0.0101	0.0382	0.0105	0.0308	-0.1028	0.0182	-0.0039	0.9270	0.0697	0.2220
LNSTUDPP	-0.0022	0.0667	-0.0007	0.2593	-0.0007	0.2540	-0.0023	0.0557	-0.0019	0.3010	0.0016	0.6850
LNRESEAR	0.0125	0.0001	0.0043	0.0039	0.0044	0.0035	0.0125	0.0000	0.0113	0.0010	-0.1594	0.0000
LNMFFCOM	0.2410	0.0001	0.1482	0.0001	0.1505	0.0001	0.2432	0.0000	0.2552	0.0000	-0.9015	0.0220
LNHESSP	-0.2086	0.0001	-0.0463	0.0001	-0.0459	0.0001	-0.2101	0.0000	-0.1722	0.0000	-0.1841	0.0000
LNTOTOP1	-0.0054	0.0069	-0.0152	0.0001	-0.0152	0.0001	-0.0055	0.0488	-0.0066	0.0000	-0.0064	0.0000
LNA_E1	-0.0017	0.0001	-0.0009	0.0075	-0.0009	0.0074	-0.0017	0.0000	-0.0011	0.0060	-0.0008	0.0360
LNAVBEDS	0.1610	0.0001	-0.0014	0.8230	-0.0014	0.8264	0.1649	0.0000	0.0970	0.0000	0.0440	0.0020
LNHEATBD	0.0523	0.0001	0.0323	0.0001	0.0324	0.0001	0.0517	0.0000	0.0245	0.0020	0.0062	0.4220
LNSITES	0.0607	0.0001	0.0320	0.0001	0.0321	0.0001	0.0597	0.0000	0.0346	0.0000	0.0042	0.6590
LNITINDX	-0.0379	0.0001	0.0026	0.5967	0.0030	0.5506	-0.0373	0.0000	-0.0279	0.0000	-0.0221	0.0030
LNPOPDEN	-0.0017	0.0879	0.0010	0.3124	0.0010	0.2876	-0.0016	0.1215	0.0170	0.0300	(dropped)	
LNHERF15	-0.0268	0.0001	-0.0250	0.0001	-0.0246	0.0001	-0.0260	0.0000	-0.0049	0.6430	(dropped)	
MERGED	0.0104	0.3023	0.0305	0.0001	0.0307	0.0001	0.0090	0.3303	0.0197	0.2430	0.1758	0.0030

YEAR45	Reference category		Reference category		-0.0141	0.0496	Reference category		Reference category		Reference category	
YEAR56	0.0354	0.0057	-0.0043	0.5965	-0.0185	0.0286	0.0339	0.0120	0.0168	0.0710	0.0039	0.6720
YEAR67	0.0228	0.0229	-0.0005	0.9429	-0.0150	0.0264	0.0229	0.0244	0.0055	0.4360	-0.0409	0.0000
YEAR78	0.0552	0.0001	0.0145	0.0437	Reference category		0.0556	0.0000	0.0280	0.0000	-0.0335	0.0010
SPEC1			Reference category		-0.0346	0.3934						
SPEC4			-0.0287	0.2181	-0.0593	0.1156						
SPEC5			-0.0913	0.0011	-0.1166	0.0027						
SPEC9			0.1414	0.0001	0.1098	0.0030						
SPEC10			0.1387	0.0001	0.1039	0.0061						
SPEC11			0.0097	0.6751	-0.0162	0.6630						
SPEC12			0.1364	0.0001	0.1006	0.0077						
SPEC13			0.1363	0.0001	0.1008	0.0079						
SPEC14			0.2699	0.0001	0.2329	0.0001						
SPEC15			0.1722	0.0001	0.1348	0.0004						
SPEC16			-0.1729	0.0001	-0.1868	0.0001						
SPEC17			0.0038	0.8785	-0.0283	0.4641						
SPEC18			-0.1683	0.0001	-0.1778	0.0001						
SPEC19			0.1108	0.0001	0.0766	0.0628						
SPEC20			-0.2235	0.0001	-0.2398	0.0001						

3. RESULTS

3.1 Model 1

The results for model 1 with outliers removed (the 'trimmed' model) are shown in Table 3. The trimming process left a sample of 892 hospital-years' worth of data (out of 951). The adjusted R-squared is around 0.71 suggesting that the full set of regressors are able to explain over two-thirds of the variation in the cost-index. There are no significant multicollinearity[4] or heteroskedasticity problems. Where explanatory variables are strongly correlated, there is sufficient statistical power to produce quite precise coefficient estimates. The model passes the RESET test ($p>0.05$).

The trimmed data coefficients are likely to be more reliable estimates than the full model estimates, as they represent the norms of more typical providers (though the signs and coefficient sizes do not differ much from the full model). In most cases, where parameter estimates are statistically significant, they also have the expected sign (Appendix 2). Exceptions include the inter-specialty transfers term (LNFCEINP), unexpected emergencies (LNEMERIN), the proportion of elderly persons (LNPROP60) and market competition (LNHERF15) all of which are associated with a lower cost index, whereas theory would suggest that they should be cost increasing.

It is not entirely clear why these variables have negative coefficients, but it is possible that transfers and emergencies are in fact measuring some other form of activity and that monopoly power is in some way confounded by merger activity. The Herfindahl index has not been updated since 1995, and the significant number of mergers that occurred in the interim would have caused changing market concentration over this period. It is possible that the effect of increasing numbers of mergers which would lead to reduced competition (and a higher Herfindahl index), may have been picked up in this variable, leading to lower cost and hence a negative coefficient. The geographic 15 mile radius boundary created by the Herfindahl index may therefore pick up more of the concentration effect than actual competition between providers for purchaser contracts. Markets are most competitive in densely populated urban areas, especially London (where in 1995, one Health Authority had 10 acute care providers within its boundaries). While we would expect this effect to drive down costs in these areas, the potential political damage caused by hospital closures has caused the Department of Health to intervene to protect many of these providers and it is likely that these results show the effect of government intervention, rather than any market phenomenon. Actually this result is consistent with some studies from the US prior to stringent cost constraints where hospitals under retrospective reimbursement in more competitive environments exhibited significantly higher costs (Robinson & Luft, 1985). More recent studies show that merging hospitals gaining market share or hospitals gaining market power with purchasers are able to negotiate higher prices (Melnick *et al*, 1992; Krishnan, 2001).

The episodes per spell term (LNEP_SPL), and the non-primary outpatient volume term (LNOPNPP), have large, significantly positive coefficients, as expected. This suggests that multi-episode spells are indicative of increased complexity and costliness. Outpatient re-attendances are associated with significant extra costs,

[4] The largest VIF was 7.87 (mean 2.35) (Chatterjee & Price, 1991).

suggesting that they do constitute an extra output and are a valid policy choice to be considered in assessing hospital performance. The HRG index term (LNHRGWT) just achieved statistical significance at the 5 percent level suggesting that the deterministic case mix adjustments applied to the dependent variable may under-compensate for more complex cases. Female patients (LPROPFEM) appear less costly than men, after other case mix adjustments. Results suggest that female patients and patients in the under 15 (LNPROP15) and over 60 (LNPROP60) age brackets are less resource consuming than other patients, all else being equal. This does seem to correspond with previous research findings that suggest that men receive more service-intensive care in order to reduce their length of stay and hence opportunity cost of lost income (Ro, 1969). Other studies have shown, that aged patients may consume fewer resources per day as a result of lower service intensity (Lave *et al*, 1972; Hornbrook & Monheit, 1985). Coefficients on the age and sex terms cannot be interpreted in isolation from the HRG index, however, since this already incorporates some age and sex adjustments and may indeed overcompensate in some HRG categories for them.

The specialisation index (LNITINDX) is significant and negative, suggesting that economies of specialisation may apply and that Trust specialisation may be associated with greater efficiency. Although many of the excluded ‘outlier’ Trusts are single specialty providers with high specialisation indices, a similar result is obtained on the full sample version of model 1. This contradicts earlier results in the cross-section modelling and probably requires further investigation. It is consistent with results from the US, however, which have suggested that non-price competitive pressures to introduce technology changes have driven specialisation over the 1980s (Luft *et al*, 1986; Farley & Hogan, 1990). International evidence suggesting better quality of care when hospitals treat higher volumes of certain medical conditions and procedures also tends to favour some specialisation (Flood *et al*, 1984; Hughes *et al*, 1987; Luft *et al*, 1987). This has important implications for the way in which Trusts are established and organised and more specific analysis is needed around the circumstances and type of care which is more efficiently provided in multi-specialty versus single specialty environments.

Inter-hospital transfer variables (LNTRANSI and LNTRANSO) do not contribute significantly to the model, and neither is the proportion of emergency admissions (LNEMERGP) a significant predictor of costs. The market forces factor (LNMMFCOM) has the expected positive sign and is highly significant, which is consistent with findings from other studies (Berry, 1970). This suggests that higher factor prices in local labour and property markets do feed through to higher patient treatment costs.

Student teaching activity (LNSTUDPP) is not significantly associated with costs, whereas research activity (LNRESEAR) has a strongly positive association with costs. It should be noted that hospital revenue for teaching and research purposes is deducted from the numerator of the dependent variable, so these coefficients represent the extent to which hospitals are over or under compensated for their academic activity. The positive sign on the research variable therefore suggests either that teaching hospitals are not fully compensated for their research associated costs, or additional income from outside the NHS is used to fund research associated increases in service costs. Previous work suggests that because of the problems in accurately measuring teaching and research output, and the fact that they are closely related, these results should be treated with caution (Söderlund, 1996).

All of the capacity variables (beds (LNAVBEDS), heated volume (LNHEATBD) and number of sites (LNSITES)) are significantly cost increasing, while scale of activity variables (number of spells (LNHESSP), first outpatient (LNTOTOP1) and first A&E attendances (LNA_E1)) are significantly cost decreasing. Together, these results confirm that hospitals with higher utilisation rates appear more efficient and thus hospitals could in principle increase efficiency ratings by increasing activity and throughput.

Recent merger (MERGED) and local area population density (LNPOPDEN) are not significant contributors to the model.

3.2 Models 2 and 3

Models 2 and 3 are estimated on specialty level data, aggregated to the whole hospital, using fixed and random effects models respectively. Both models failed the RESET test and suffered significant heteroskedasticity problems. This is perhaps indicative of the problems alluded to earlier regarding the validity of applying assumptions regarding hospital level behaviour to the level of individual specialties. Even if hospital level effects hold on average, they are likely to be of varying importance for different specialties, hence the poorer fit. Since many of the specialty effects are strongly correlated with other regressors (X and Z), parameter estimates differ significantly to those of model 1. A Hausman test between the fixed and random effects models could not be computed because the variance-covariance matrix could not be inverted (Hausman, 1978). Given its theoretical limitations, data availability and classification problems, and poor model fit, it is hard to find any advantage that these models offer over the hospital level formulation of model 1. Consequently for the purposes of this paper no further use has been made of specialty level formulations.

3.3 Model 4

Model 4 differs from model 1 by assuming an asymmetric (half-normal) error term such that a one-sided set of residuals are generated relative to an efficiency frontier (as opposed to an OLS regression line). The functional form is still log-linear, however, and the parameter estimates generated are, as expected, virtually identical to those of model 1. The stochastic frontier model does offer the advantage that it takes account of random noise, whilst still comparing Trusts to best, rather than average practice. It will therefore likely produce higher efficiency scores than an OLS model (and less stringent targets if used in a target-setting process).

3.4 Models 5 and 6

Models 5 and 6 are specified as longitudinal models at the whole hospital level using random and fixed effects models respectively, where provider specific effects are used to capture efficiency, rather than simply taking an average of residuals.

Longitudinal models in general have two advantages over cross-sectional ones. Firstly, they will control for some unmeasurable, but probably important, time invariant characteristics of providers, and consequently parameter estimates may be less susceptible to omitted variable bias. Secondly, they produce a standard error for the efficiency index of each Trust, and thus allow comparison of not only the point estimates, but also the precision of efficiency rankings.

Unsurprisingly, specifying longitudinal models in this paper proves important in terms of a significant Breusch and Pagan Lagrange Multiplier test (Greene, 1993) suggesting that provider-specific effects are important to capture in a longitudinal model since there is heterogeneity between Trusts and a random or fixed effect model would be appropriate. A significant Hausman test (Hausman, 1978) suggests that the fixed effects parameter estimates are probably more appropriate since they are consistent. However, neither model passed the RESET test. The signs and significance levels of some of the parameter estimates in models 5 and 6 differ from those in model 1 (the pooled model) and some have signs inconsistent with theory. Some variables were only measured once during the study period (population density (LNPOPDEN) and Herfindahl index (LNHERF15)) and thus drop out in the fixed effects model.

Standard errors for both fixed and random effects models were calculated as well as confidence limits for the random effects index[5]. A total of 48 Trusts (approximately one fifth) have a 95 percent confidence limit lower bound that falls above 1, the national average. It is desirable that confidence intervals be calculated in any efficiency analysis, as much of the variability in efficiency scores or rankings may be due to nothing more than sampling error (Jensen, 2000).

The fixed effects longitudinal model presents an opportunity for an interesting variant on cost benchmarking. Since it essentially measures effects as changes relative to other years for the same Trust, it allows the construction of ‘efficiency indices’ from the average residual obtained for each Trust. Although beyond the scope of this paper, this would seem to be an important addition to the Department of Health’s benchmarking armoury, and would allow comparison of the efficiency improvements between Trusts over time. It may of course be more difficult to model since efficiency rankings may be changing over time.

3.5 Efficiency scores

Trusts were ranked against one another on the three cost indices (CCI, 2CCI and 3CCI) from the pooled data for all four years. Three additional adjusted indices were calculated from whole hospital level data – those derived from models 4, 5 and 6. The basic descriptive statistics for these indices are shown in Table 4.

[5] A similar statistic can of course be calculated for the OLS derived indices since they are calculated from the mean of up to four period residuals, and the necessary parameters (sd and n) for the calculation of a standard error are thus present.

Table 4: Descriptive Statistics For The Efficiency Scores

	CCI	2CCI	3CCI	Frontier Index	Random Effects Index	Fixed Effects Index
Mean	1.0241	1.0001	1.0001	1.0000	1.0025	1.0000
N	245	245	245	239	239	239
Std. Deviation	0.2539	0.1594	0.0924	0.0327	0.0717	0.2490

Sample sizes differ slightly because the panel and maximum likelihood estimated models (4, 5 and 6) could not incorporate Trusts which appeared for only one time period. The degree of dispersion in cost indices should decrease as a greater proportion of variation is explained by the included regressors. This pattern is maintained across the OLS derived indices with a decrease in standard deviation from 0.25 (CCI) to 0.16 (2CCI) to 0.09 (3CCI). The degree of dispersion in the fixed effect index, on the other hand, is very large. This is probably because the Trust level effects capture much of the variance in place of the actual explanatory variables (X and Z vectors). This does not happen with the random effects index. This is probably due to the relatively small number of time periods relative to cross-sections, and the different ways in which random and fixed effects designs treat these two variance components. The frontier model, on the other hand, produces a set of residuals that are clustered very closely. This is probably because the half-normal error term used better fits the skew of the residuals in the data. Overall, however, increased statistical adjustments reduce efficiency variation between providers considerably, such that there is relatively little difference between providers in terms of fully adjusted (short-run) efficiency scores. This would suggest that the potential for savings by bringing poorly performing hospitals up to the level of the best ones (should this be possible) is modest, at least in the short-run. Variation in long run efficiency is greater, however, suggesting that there is still room for efficiency enhancement by optimising capacity and activity levels in the longer run.

Rank correlation coefficients between each of the efficiency indices have been calculated to facilitate further comparison as shown in Table 5.

TABLE 5: Spearman Rank Correlation Coefficients

	CCI	2CCI	3CCI	Frontier Index	Random Effects Index	Fixed Effects Index
CCI	1.00					
2CCI	0.72	1.00				
3CCI	0.63	0.77	1.00			
Frontier Index	0.62	0.76	0.97	1.00		
Random Effects Index	0.65	0.77	0.91	0.92	1.00	
Fixed Effects Index	0.61	0.32	0.46	0.46	0.57	1.00

Note: All correlations are significant at the .01 level

Interestingly, the correlation between the CCI and each of the adjusted indices is greater than 0.6 in all cases. Likewise, there is strong concurrence between all of the adjusted indices except for the fixed effects index. The 2CCI is a partially adjusted form of the 3CCI and its closer correlation with the CCI is thus not surprising. The remaining three fully adjusted indices (3CCI, random effects and frontier indices) are very closely correlated. The stochastic frontier index is particularly closely correlated with the 3CCI, which is not surprising given that the estimation technique for stochastic frontiers essentially involves shifting the intercept in accordance with the asymmetric error distribution.

4. CONCLUSIONS, POLICY IMPLICATIONS AND FUTURE RESEARCH

This paper aimed to explore various panel data modelling approaches to hospital benchmarking within the Department of Health framework of statistically adjusted cost indices. From this paper, the following conclusions can be drawn:

1. The pooling of multiple years of data is desirable in that it removes the effect of one-off data errors and provides a sample size sufficient to adequately assess the importance of most contending explanatory variables. This advantage should not however preclude single-year cross-sectional comparisons. The disadvantage to pooling the data is not being able to examine any shifts in Trust efficiency over the time periods. It would seem advisable from a target-setting point of view (whichever approach is taken) not to redesign the adjustment model every year.

2. Specialty level models appear to be inferior to whole hospital estimations by both theoretical and empirical criteria. If, at some future stage, specialties came to operate as free-standing business units within hospitals and data were collected accordingly, then specialty level efficiency comparisons might be reassessed.

3. Stochastic frontier models have the advantage over OLS and other deterministic models of taking account of random variation. For target-setting purposes the efficiency scores will necessarily be higher (and targets less stringent) compared to ordinary regression.

4. Estimating provider effects in panel models has the advantage that it automatically produces a standard error which allows estimation of a confidence limit around each efficiency score. It is important that confidence intervals be estimated for efficiency scores as differences in rankings are often statistically insignificant.

5. Fixed effects longitudinal models also have the advantage over cross-sectional (or pooled) models in that they may take account of some important time invariant Trust characteristics and parameter estimates may therefore be less susceptible to omitted variable bias.

6. Use of the residuals from longitudinal fixed-effect models might be useful in constructing efficiency indices to examine Trust efficiency improvement over time.

7. Significant data deficiencies still exist, and reducing these would contribute more to better modelling than further experimentation with alternate specifications and estimation techniques. In particular, data on research and teaching outputs should be collected on an annual basis. In the longer term, efforts to collect quality of care and outcome indicators would be important.

8. Although there were not large differences between the full sample model, and one estimated with influential observations removed, the current quality of NHS

data suggest that both forms should be used. If they differ significantly, the outlier-removed model is probably preferable.

This paper differs from much of the academic literature in this area by assuming an explicitly normative framework. Within this framework, the focus shifts from trying to discover how hospitals work to explaining why NHS hospitals don't work as expected. The emphasis of the paper is to work within the Department of Health's benchmarking framework to explore ways of monitoring poorer performing Trusts. On both econometric and theoretical grounds the estimation of total cost functions is usually preferred to average cost functions (or in this case the cost index) for efficiency analyses (Vitaliano, 1987). Given the multi-output nature of hospitals, more flexible functional forms may lead to more accurate inferences about coefficients (and hence efficiency estimates) (Vita, 1990). This will allow a more defensible interpretation of the model residuals as deviations from the cost frontier, representing inefficiency. It may therefore be prudent for the Department of Health to consider embarking on the more conventional efficiency analysis path in future with respect to dependent variable and functional form choice as this is less objectionable in terms of economic theory.

The degree of "fully adjusted" variation between hospitals in terms of efficiency after adjusting for factors exogenous to managerial control, is not great. When the short-run adjustments are removed, however, variation increases, suggesting that there is scope for efficiency enhancement by optimising capacity and activity levels in the longer run. This is a fairly obvious, but often overlooked, fact - that hospitals are more efficient when full, and reduction of surplus capacity may be the most obvious way to achieve efficiency gains. The crucial flaw in this concept of efficiency is that it omits important outputs such as waiting time, responsiveness and quality of care. Hence greater activity levels may have adverse consequences such as increased hospital infection rates, it may crowd out other vital outputs such as holding sufficient capacity for stochastic emergency demand, and it may possibly sideline quality of patient care (Bagust *et al*, 1999). Emergency capacity is critical since hospitals running at high occupancy rates struggle to efficiently deliver emergency care and suffer periodic bed crises such as the winter bed pressures. Bed shortages will also have a deleterious effect on patient services.

Within the current Department of Health benchmarking framework explored in this paper, caveats remain. Without more information on quality, one cannot say that high or low costs are indicators of efficiency, nor that these providers produce effective or high quality services. This will remain a caveat in efficiency rankings based on routinely gathered data and information on quality and health outcomes should be used to moderate judgements of Trust efficiency.

Ultimately, because Trusts are not truly independent economic agents and much of their performance depends on historical and current decisions at a central level, it would seem that a collaborative response to change both Trust and exogenous environmental factors would be most useful.

5. REFERENCES

Bagust, A., Place, M. & Posnett, J.W. (1999) Dynamics of bed use in accommodating emergency admissions: Stochastic simulation model, *British Medical Journal*, 319: 155-8.

Belsey, D.A., Kuh, E. & Welsch, R.E. (1980) *Regression diagnostics*, New York: John Wiley and Sons, Inc.

Berry, R.E. (1970) Product heterogeneity and hospital cost analysis, *Inquiry*, 7(1): 67-75.

Braeutigam, R. & Daughety, A. (1983) On the estimation of returns to scale using variable cost functions, *Economics Letters*, 11: 25-31.

Breyer F. (1987) The specification of a hospital cost function: A comment on the recent literature, Journal of Health Economics, 6: 147-57.

Butler, J.R.G. (1995) *Hospital cost analysis*, Kluwer Academic Publishers: Dordrecht.

Chatterjee, S. & Price, B. (1991) *Regression analysis by example 2nd edition*, New York: John Wiley and Sons, Inc.

Cook, R.D. & Weisberg, S. (1983) Diagnostics for heteroscedasticity in regression, *Biometrika*, 70: 1-10.

Culyer, A.J., Wiseman, J., Drummond, M.F. & West, P.A. (1978) What accounts for the higher costs of teaching hospitals? *Social and Economic Administration*, 12(1): 20-30.

Farley, D.E. (1989) Measuring casemix specialization and the concentration of diagnoses in hospitals using information theory, *Journal of Health Economics*, 8: 185-207.

Farley, D.E. & Hogan, C. (1990) Case-mix specialization in the market for hospital services, *Health Services Research*, 25: 757-83.

Flood, A.B., Scott, W.R. & Ewy, W. (1984) Does practice make perfect? Parts I and II, *Medical Care*, 22(2): 98-124.

Freund, R.J. & Littell, R.C. (1991) *SAS system for regression*, Cary, NC: SAS Institute.

Friedlander, A. & Spady, R. (1981) *Freight transport regulation: Equity efficiency, and competition in the rail and trucking industries*, Cambridge, MA: MIT Press.

Giuffrida, A. & Gravelle, H. (1998) *Hospital benchmarking: Comments*, National Primary Care Research and Development Centre, Centre for Health Economics, University of York.

Greene,W.H.(1993) *Econometric analysis*, New York: MacMillan.

Hausman, J. (1978) Specification tests in econometrics, *Econometrica*, 46: 1251-71.

Hornbrook, M.C. & Monheit, A.C. (1985) The contribution of case-mix severity to the hospital cost-output relation, *Inquiry*, 22(3): 259-71.

Hughes, R.G., Hunt, S.S. & Maerki, S.C. (1987) Effects of surgeon volume and hospital volume on quality of care in hospitals, *Medical Care*, 25(6): 489-503.

Jensen, U. (2000) Is it efficient to analyse efficiency rankings? *Empirical Economics*, 25: 189-208.

Krishnan, R. (2001) Market restructuring and pricing in the hospital industry, *Journal of Health Economics*, 20:213-37.

Lave, J.R., Lave, L.B. & Silverman, L.P. (1972) Hospital cost estimation controlling for case-mix, *Applied Economics*, 4: 165-80.

Luft, H.S., Robinson, J.C., Garnick, D.W., Maerki, S.C. & McPhee, S.J. (1986) The role of specialized clinical services in competition among hospitals, *Inquiry*, 23: 83-94.

Luft, H.S., Hunt, S.S. & Maerki, S.C. (1987) The volume-outcome relationship: Practice makes perfect or selective referral patterns? *Health Service Research*, 22(2):157-82.

Melnick, G.A., Zwanziger, J., Bamezai, A. & Pattison, R. (1992) The effects of market structure and bargaining position on hospital prices, *Journal of Health Economics*, 11: 217-33.

Noble, G. & Lawrence, R. (2000) *NHS performance and taxpayers' money: A view from HM Treasury*, in Appleby, J. & Harrison, A., Health Care UK, King's Fund Publishing: London.

Ramsey, J.B. (1969) Tests for specification error in classical linear least squares regression analysis, *Journal of the Royal Statistical Society*, B31: 250-71.

Ro, K.K. (1969) Patient characteristics, hospital characteristics and hospital use, *Medical Care*, 7(4): 295-312.

Robinson, J.C. & Luft, H.S. (1985) The impact of hospital market structure on patient volume, average length of stay, and the cost of care, *Journal of Health Economics*, 4: 333-56.

Söderlund, N., Milne, R., Gray, A., & Raftery, J. (1995) Differences in hospital casemix, and the relationship between casemix and hospital costs, *Journal of Public Health and Medicine*, 17: 25-32.

Söderlund, N. (1996) *The Impact of Medical Research and Development Activity on NHS Hospital Treatment Costs - 1991/2 to 1993/4 - A second report for the R&D Directorate, NHS Executive*, Unpublished monograph.

Söderlund, N., Gray, A., Milne, R., & Raftery, J. (1996) Case mix measurement in English hospitals: An evaluation of five methods for predicting resource use, *Journal of Health Services Research and Policy,* 1: 10-9.

Söderlund, N., Csaba, I., Gray, A., Milne, R. & Raftery, J. (1997) The impact of the NHS reforms on English hospital productivity: An analysis of the first 3 years, *British Medical Journal*, 315: 1126-9.

Söderlund, N. & van der Merwe, R. (1999) *Hospital benchmarking analysis and the derivation of cost indices*, Centre for Health Economics Discussion Paper 174: University of York.

Tatchell, P.M. (1980) *Measuring hospital output: A review*, Australian National University Press: Canberra.

Vita, M.G. (1990) Exploring hospital production relationships with flexible functional forms, *Journal of Health Economics*, 9: 1-21.

Vitaliano, D.F. (1987) On the estimation of hospital cost functions, *Journal of Health Economics*, 6: 305-18.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, 48: 817-38.

6. APPENDICES

Appendix 1
Aggregated Specialties

Minor specialties, supra-district and supra-regional specialties were allocated to appropriate main specialty categories and allocated the following codes.

SPEC1	Paediatrics, paediatric neurology and paediatric surgery
SPEC4	General medicine, endocrinology, clinical physiology, clinical pharmacology, audiological medicine, clinical genetics, clinical cytogenetics and molecular genetics, palliative medicine, nuclear medicine, clinical neuro-physiology, dental medicine specialties, community medicine, occupational medicine, dermatology, infectious diseases, genito-urinary, nephrology, thoracic medicine, clinical immunology, rehabilitation medicine, geriatrics, general practice (other then maternity), neurology, haematology, medical oncology, rheumatology, gastroenterology
SPEC5	Cardiology
SPEC9	General surgery
SPEC10	Urology
SPEC11	Orthopaedics
SPEC12	ENT
SPEC13	Ophthalmology
SPEC14	Gynaecology
SPEC15	Dental surgical specialties
SPEC16	Neurosurgery
SPEC17	Plastic surgery
SPEC18	Cardiothoracic surgery
SPEC19	Obstetrics
SPEC20	Mental illness and mental handicap
SPEC22	Miscellaneous other specialties (excluded from analysis)

Appendix 2
Explanatory variables used

COSTINDX *Cost index*
The cost index was calculated at the hospital level as:

$$CCI_i = \frac{C_i}{[(\underline{IC}*IP_i*HI_i)/(\underline{HI}*\underline{IP})] + \Sigma_j OP_{ij}*\underline{OC}_j/\underline{OP}_j + AE_i*\underline{AC}/\underline{AE}}$$

where:
CCI_i = Cost index for hospital i
C_i = Cost incurred by hospital i for inpatient, outpatient and A&E care
$\underline{IC}$ = Total costs incurred for inpatient spells for all acute hospitals
IP_i = Number of inpatient spells in hospital i
HI_i = HRG case mix index for hospital i
$\underline{HI}$ = Average HRG case mix index for all spells treated in the study hospitals
$\underline{IP}$ = Total number of inpatient spells in all study hospitals
OP_{ij} = Total first outpatient attendances in hospital i in specialty j
$\underline{OP}_j$ = Total first outpatient attendances for all study hospitals in specialty j
$\underline{OC}_j$ = Total cost of outpatient attendances for all study hospitals in specialty j
AE_i = Total first A&E attendances in hospital i
$\underline{AE}$ = Total first A&E attendances in all study hospitals
$\underline{AC}$ = Total cost of A&E in all study hospitals

Costs include a notional capital cost, estimated as 6 percent of the net asset value of the Trust attributable to inpatient, outpatient and A&E services delivered to NHS patients. Service Increment for Teaching and Research (SIFTR) revenue, and revenue from private patients were excluded from the total cost figure in an attempt to obtain a pure measure of NHS patient care costs. Net costs obtained may thus be biased to the extent that these revenues over or under compensate Trusts for excluded activities.

TRANSIPP *Proportion of spells that involve a transfer in from another hospital*
Hypothesised sign – positive
Transfers in to a hospital are likely to represent difficult or problem cases referred from less capable institutions and are thus likely to be cost increasing.

TRANSOPP *Proportion of spells that end in a transfer to another hospital*
Hypothesised sign – negative
It is assumed that transfers from a hospital represent an inability to meet the treatment needs of a given case. Transfers out of a hospital are likely to represent incomplete treatment of cases and are thus likely to be cost decreasing.

EMERGPP *Proportion of spells that involve an emergency admission*
Hypothesised sign – positive
Large fluctuations in levels of emergency admissions imply that more fixed capacity has to be retained for a given average level of activity, and consequently, costs should increase. This variable measures whether, diagnostic case mix and other factors being equal, emergencies will be more costly than elective admissions because of the implied threat of serious adverse outcome.

FCEINPP *Proportion of spells involving a transfer in from another specialty*
Hypothesised sign – positive

Measurement of the volume of inpatient care performed by NHS acute hospitals has been the finished consultant episode (FCE). During a single hospital admission, however, multiple FCEs might occur as a result of transfers within or between specialties. The inpatient spell, or set of episodes constituting a single admission, thus serves as a slightly higher level of aggregation of inpatient activity. Although the FCE has been extensively criticised, it is argued that spells also fail to fully capture total inpatient activity. Both variables have therefore been captured in the model. Spells requiring inter-specialty transfers are likely to be more complex and costly than those which can be fully treated within a specialty. Given the existing adjustment for episodes per spell, this variable captures the additional effect of inter-specialty transfers over and above the average multiple FCE.

EP_SPELL *Average NHS inpatient episodes per NHS inpatient spell*
Hypothesised sign - positive
Although the model used treatment spells (whole admissions) as the measure of volume of inpatient activity for Trusts, the episodes variable incorporates the fact that volume of inpatient activity is represented by two variables (the spell and the episode). It is argued that the true unitary measure of volume of inpatient activity probably lies somewhere between the spell and the episode. This variable differs from the above one in that it encompasses all multi-episode spells, and not just those between specialties. If significantly positive, it would indicate that even multiple episode spells within a specialty were more costly that single episode equivalents.

OPNPP *Non-primary outpatient attendances per inpatient spell*
Hypothesised sign – positive
The basic unit of outpatient activity is assumed to consist of first, rather than follow-up outpatient attendances, based on information which suggests that many outpatient attendances occur because of a failure to complete treatment during the first attendance. Since follow-up attendances may in some instances constitute genuine additional health care output, this variable has been included as a regressor.

EMERINDX *Standardised index of unexpected emergency admissions / total emergencies*
Hypothesised sign – positive
This variable reflects additional costs associated with coping with unpredictable demand. The variable is calculated as the 12-month sum of the absolute value of residuals from a simple model of emergency admissions[6] standardised to give an index with a national average of one.

HRGWTNHS *HRG case mix index*
Hypothesised sign – positive
Healthcare Resource Groups (HRGs) were taken to be the best available categorisation system for inpatient case mix in the hospitals studied. In order to estimate a case mix index for a hospital, all cases were allocated to an HRG category, and a weight representing the expected cost of that category attached accordingly. The average cost weight for all spells treated over a year formed the scalar case mix index for that hospital. The national average case weight was set to equal 100, and case mix

[6] A single regression was run for the whole sample using provider specific dummies with each month's emergency demand set as a function of the previous months' demand plus a monthly dummy.

indices above 100 thus represent hospitals that have treated a more complex than average mix of cases.

While the use of a single index to represent case type variation across a possible 534 categories is somewhat reductionist, previous studies have shown that case mix is a powerful predictor of hospital costs (Tatchell, 1980; Butler, 1995). When more comprehensive methods of incorporating HRG mix were used, such as principle components, a better model fit was achieved, but at considerable expense to ease of interpretation.

The case mix index remains an incomplete measure of expected patient costs for a number of reasons. Firstly, it applies only to inpatient costs, and thus excludes case-type variation amongst outpatient and A&E attenders. Secondly, while some age and gender splits occur within HRGs, these are not universal, and one would expect some residual effect of age and gender on hospital costs after HRG-based adjustments. Thirdly, aspects of patient case mix, especially those related to severity of illness, are generally not completely captured by diagnosis-based measures such as HRGs. Therefore proxy variables such as the number of emergencies and whether or not cases have been transferred in from elsewhere have been used to capture some of this severity effect.

PROP15U *Proportion of patients under 15 years of age*
Hypothesised sign – positive
This variable measures whether social expectations may force Trusts to expend more resources on younger patients, diagnosis and other factors being equal (Söderlund *et al*, 1995).

PROP60P *Proportion of patients 60 years or older*
Hypothesised sign – positive
Elderly patients are likely to have more complex care needs, and these may not be captured entirely by HRGs, which have only limited age sensitivity.

PROPFEM *Proportion of female patients*
Hypothesised sign – uncertain
This variable was inserted to capture any gender-specific differences in resource need, other case mix factors being equal.

STUDENPP *Medical student whole time teaching equivalents per inpatient spell*
Hypothesised sign - positive
Student teaching typically constitutes the major “academic activity” cost driver in academic hospitals. Student whole-time equivalents (WTEs) are used as an indicator of the amount of teaching done by a hospital. The data relates to two NHS Executive Surveys from 1992/3 and 1997/8. Figures for intervening years were calculated by means of a weighted moving average.

RESEARPC *Percentage of total revenue spent on research*
Hypothesised sign – uncertain
Teaching and research activities constitute important secondary outputs of NHS hospitals. They have a well-documented positive impact on hospital costs (Culyer *et al*, 1978). In this case, however, the compensation which hospitals receive for

teaching and research (SIFTR) has already been deducted from the cost component of the dependent variable *a priori* (the cost index). These terms thus capture the extent to which hospitals incur research costs over and above the compensation they receive for these activities. Alternatively, assuming that Trusts are fairly compensated for teaching and research outputs produced, these terms capture inefficiencies associated with teaching and research. The data relates to a survey conducted by the NHS Executive in 1994.

MFF_COMB *Market forces factor*
Hypothesised sign – positive
Market prices for inputs including land, buildings and labour differ between Trusts because of their geographic location. This represents an unavoidable influence on hospital costs. Component price indices are weighted according to their proportional contribution nationally in constructing the index. The index was calculated annually using the component market forces factors and weightings supplied by the Department of Health. In reality, very little change occurs in this index year on year.

POPDENS *Population density – average head of population per square mile*
Hypothesised sign – uncertain
More densely populated urban areas typically would be expected to have greater stability of patient demand and hence be able to maintain occupancy levels, improve cost-effectiveness, and lower costs. They would also be able to discharge earlier to community based care, which would be more feasible in densely populated areas.

HERF15 *15 mile radius Herfindahl Competition index*
Hypothesised sign - Positive
The greater the number of local providers, the lower we would expect costs to be as a result of competitive pressures. The index has a maximum value of 1 where there are no other acute providers within a 15-mile radius. Indices were calculated for previous research (Söderlund *et al*, 1997). The index used here is the 1995/6 index and has not been updated for subsequent years.

HESSPNHS *Total inpatient spells*

TOTOP1 *Total first outpatient attendances*
It was felt that first outpatient and A&E attendances better represented health care output in these areas than did total number of attendances. In the case of A&E, subsequent attendances are likely to be for minor interventions such as removal of sutures or change of dressings and would therefore not incur significant costs.

A_E1 *Total first accident and emergency (A&E) attendances*
Hypothesised signs – negative
After adjustment for the levels of fixed inputs used by a Trust, increased volume of activity is expected to lower average costs (or increase efficiency).

AVBEDS *Average available beds*
Hypothesised sign – positive
Average bed numbers may be considered fixed in the short-run. While hospital managers do have some control over the size and capacity of their institution, it is expected that there will be some reluctance to radically alter capacity. Decreasing

hospital capacity might be particularly difficult because of public opposition and implied job loss. The average number of beds in a hospital is thus included to reflect an inability to alter capacity in the short run. It is however expected that an increase in beds (capital stock) will increase fixed costs.

HEATBED *Heated volume per bed*
Hypothesised sign – positive
This variable was included to capture inefficiencies in how hospital buildings were used to create treatment capacity (represented by beds). A large amount of heated volume per bed was assumed to represent less efficient use of capital, and thus increase the cost index.

SITES50B *Number of sites with more than 50 beds*
Hypothesised sign – positive
Trusts that are located on a number of sites, rather than concentrated in one location, are likely to suffer from duplication of certain capital and staff inputs, as well as incurring communication and management difficulties, thus increasing costs. The number of major sites with more than 50 beds was chosen to exclude sites that were simply isolated accommodation, chronic care or outpatient facilities.

ITINDX *Information Theory specialisation index*
Hypothesised sign – positive if scope economies exist; negative if specialisation economies exist.
Single specialty hospitals are likely to draw patients from further afield, and have greater short-term variation in demand for services because of the lack of cross-specialty compensation effects. Economies of specialisation on the other hand, might occur where relatively under-utilised, specialised fixed resources are centralised in one institution, rather than spread over many. This can be examined through the inclusion of an Information Theory Index which calculates the degree to which the proportions of different case-types (HRGs) in a hospital differ from the national average proportions of case-types. The formula used for derivation of the Information Theory Index as calculated by Farley (Farley, 1989; Farley & Hogan, 1990) is given below:

$$\mathrm{ITI}_h = \Sigma_i \mathrm{P}_{ih} \log (\mathrm{P}_{ih} / \pi_i)$$

where:
ITI_h = case mix specialisation index for hospital *h*
P_{ih} = proportion of cases in hospital *h* that fall into HRG *i*
π_i = proportion of all hospitals' caseload constituted by HRG *i*
An increased IT index indicates a relatively more specialised hospital (i.e. one with a narrower scope of activities) which one would expect to be of higher cost. General hospitals typically have an IT index of between 0.2 and 0.5, whereas this may increase to up to 2.5 in a highly specialised, single discipline, hospital.
Despite the fact that they both use HRGs in their construction, the specialisation index and the HRG case-mix index are fundamentally different. The former simply captures the range of different types of cases treated, whereas the latter captures the average resource intensity of cases.

MERGED *Dummy variable indicating the Trust has been part of a merger*
Hypothesised sign – negative

The variable takes a value of zero for all observations except for Trusts that are in their first year post merger. The underlying hypothesis is that cost synergy savings are made fairly immediately after a merger.