IMA Journal of Management Mathematics Advance Access published April 12, 2005

IMA Journal of Management Mathematics Page 1 of 13
doi:10.1093/imaman/dpi015

Length of stay as a performance indicator: robust statistical methodology

ELENA KULINSKAYA†
Statistical Advisory Service, Room G06, Sir Alexander Fleming Building, South Kensington Campus, Imperial College, London SW7 2AZ, UK

DIANA KORNBROT
Faculty of Health and Human Sciences, University of Hertfordshire, Hatfield ALIO 9AB, UK

AND

HAIYAN GAO
Intensive Care National Audit and Research Centre, Tavistock House, Tavistock Square, London WC1H 9HR, UK

Length of stay (LOS) is an important performance indicator for costing and hospital management and a key measure of efficiency of NHS. However, LOS is difficult to analyse because its statistical distribution is non-normal and LOS data habitually have many outliers. Furthermore, the usefulness of LOS for improving NHS performance is undermined because no adjustments are made for some key factors. This paper addresses both these problems. Health episodes statistics data from the UK NHS for 1997/98, and 1998/99 are analysed to investigate the effects of five key variables: admission method, discharge destination, provider (hospital) type, speciality and NHS region. All are found to influence LOS. The effects of some factors are substantial, and were not previously known, and so are not included in planned future NHS performance measures, e.g. LOS is at least 25% longer for patients transferred from other hospitals rather than admitted as an emergency; and LOS for patients discharged to private institutions is more than twice that for patients discharged to NHS institutions or their own home. The problem of finding the most appropriate statistical analysis for data of the LOS type is addressed by comparing standard general linear model methods with an advanced robust method called truncated maximum likelihood (TML). The TML methods are shown to have several advantages over standard methods, in terms of model fit and accuracy of parameter estimation. Implications of these findings for future use of LOS are considered.

Keywords: hospital performance indicators; adjustment; HRG classification; robust methods; general linear model; truncated maximum likelihood methods.

1. Introduction

Length of stay in hospital (LOS) is a crucial variable for the quality of life of all patients and their families. Furthermore, it is the single most important component in the consumption of hospital resources (Lave & Leinhard, 1976). It is also very important for hospital planning since it is a direct determinant of the number of beds to be provided. Moreover, LOS is a frequent point of comparison between patients, hospitals and countries. Hence, LOS is a key performance indicator for hospital management and a key measure of efficiency of the NHS. Consequently, understanding the factors that influence LOS is a very important problem.

This paper has two purposes. The first purpose is to elucidate what factors influence LOS, over and above, age, sex and Health Resource Groups (HRG) classification (which are always included in any

†Corresponding author. Email: E.Kulinskaya@imperial.ac.uk

estimate). The extra factors considered are: admission method, discharge destination, provider hospital type, NHS region and speciality (acute or geriatric). The second purpose is to determine the most accurate and consistent statistical method for analysing LOS and other similar data. The data analysed is from the UK NHS for the years 1997–1998 and 1998–1999 for Chapter A of the HRG classification—The Nervous System. Consistency of results from year to year is obviously an important criterion for an effective statistical analysis and is used in our evaluations of the methods.

We first discuss briefly the need for robust methods of statistical analysis for LOS data. Then, we summarise techniques for classifying patients according to health resource needs, together with the associated use of LOS for costing and as a high-level performance indicator. A methodological section describes the robust statistical methods used here. The results section applies these methods to the NHS data. The discussion section considers the implications of these results, both for key factors determining LOS and for the best methods of statistical analysis.

1.1 *The need for robust statistical methods*

The statistical analysis of LOS is bedevilled by the presence of outliers, i.e. unusually long stays that are very variable in their length and in their occurrence, e.g. one year a particular trust may have a maximum LOS of 60 days, another year 90 days, a third year no stays over 30 days. Including all data may have unfortunate consequences for the 95% of patients who are *not* outliers, because the outliers may have disproportionate effects on the means of subgroups of patients. The present paper considers more efficient methods of removing outliers, together with alternative robust methods of analysing data.

Robust statistical methods address the problem of non-normal data in a variety of ways, all of which down weight outliers. It is a common misconception that robust methods that rely on estimates based on only a proportion of the data, say 90%, ignore outliers. This is not the case for any of the theoretically sound robust methods, Huber (1981), Hampel *et al*. (1986), Staudte & Sheather (1990). Rather these methods predict the behaviour of the outliers from the behaviour of the more central values. This ensures that the central data and extreme, but non-outlier, data are accurately estimated. The distinction between outliers and extreme values is important. Extreme values are simply the highest values in a data set, and so of course are always present. Outliers are values that are higher than would be predicted from parameters, such as the mean and standard deviation, of some population distribution. The underlying assumption is that the actual patients come mostly from a predominant distribution that is normal, or can be transformed to normality, but that outliers come from a different population. The log transformation on LOS data is known to generate a near normal distribution, consequently more sophisticated methods such as Box–Cox transformations are unnecessary (Kulinskaya *et al.*, 2001). Transformations will not remove the outliers, which is why robust methods are so important. The aim of the statistical analysis is twofold: to estimate the parameters of the predominant distribution; and to estimate what proportion of patients need to be considered separately as outliers. It is not suggested that the outliers be ignored in any costing calculations! Far from it. Rather it is suggested that costs should be based on both the predominant population, accurately estimated by robust methods and on contingency plans for a known proportion of outliers.

1.2 *LOS, patient classification systems, costing and efficiency*

Obviously, healthcare costs and LOS depend on the medical procedures used to treat the patient. Patient classification systems group together patients with similar resource needs, within larger organ-body system classifications such as ‘the nervous system’. Mean LOS for a resource group is used for two different audit and planning purposes: first for estimating costs and second as a high-level performance indicator.

The patient classification system used in the UK is termed Healthcare Resource Groups (HRGs). The current *Version 3.1 HRGs* was produced in 1997 (National Casemix Office, 1997). It includes 572 HRGs subdivided into 18 'chapters' corresponding to main organ-body systems. At present, the HRG classification is undergoing major development (with version 4.0 due in April 2005), in preparation for its new role for casemix adjusted cost and volume commissioning from 05/06. See http://www.nhsia.nhs.uk/ casemix/pages/admitted_hrgs.asp.

1.3 *LOS and costing*

Once patient classification systems have been used to assign patients to resource groups, costings can be based on the proportion of patients in each resource group. This is known as casemix-based costing. The development of such systems is taking place around the globe (Fetter, 1991). LOS is also a direct target in strategies aimed at the control of hospital costs (Deyo *et al*., 1986; Muschlin *et al*., 1991).

1.4 *LOS as a high-level performance indicator*

Performance assessment is central to the drive for higher quality standards within the NHS. *The high-level performance indicator set* was first published by NHSE in June 1999 (NHSE, 1999). One of the goals of its development was its use as the basis for benchmarking performance of NHS organisations. Casemix-adjusted LOS is one of the key benchmark indicators. Further publications of *NHS performance Indicators* (NHSE, 2000, 2002) contain an expanded set of performance indicators. The data are to be standardised (indirectly) only by age, HRG and (from 2002) sex. However, there are no proposals for adjustment of risk, or any other factors, in spite of the fact that current classification and performance measurement systems do not fully exploit available information on admission, severity of illness, Roe *et al*. (1996), Briggs & Gray (2000), as well as co-morbidities, Roe *et al*. (1998), or provider type and speciality, and socio-economic characteristics. The present study will provide evidence as to whether other factors should be included. It will also demonstrate how appropriate statistical techniques for modelling LOS with the help of covariates can fully exploit available data. These techniques would be useful both for refining current patient classification systems and for improving performance measurement in the health services.

1.5 *Robust statistical methods for analysis of LOS data*

The use of LOS both for costing and as a high-level performance indicator highlights the need for appropriate statistical analysis. The characteristics of LOS data have a strong effect on the nature of the most appropriate analysis. Here, we first recapitulate the properties of LOS distributions and then describe robust methods for dealing with such data.

1.5.1 *Statistical properties of LOS and traditional analyses.* The statistical analysis of LOS is made difficult for several reasons. LOS distributions in each HRG are asymmetric and have different variances in each HRG. Furthermore, LOS data typically contain outliers. A few outliers can completely distort both LOS means and comparisons based on them. This difficulty is exacerbated by the different amounts of variability in the different HRG groups. Standard analyses confront these problems by first excluding outliers using the 1.5 times interquartile range rule, and *then*, at best, log transforming the data. Analysis is then performed using the standard general linear model methods assuming that all HRGs have the same variance.

1.5.2 *Review of robust approaches.* Robust regression procedures were first proposed in the 70s and 80s (Huber, 1981; Hampel, 1978; Hampel *et al*., 1986) Key references include: Hampel *et al*., 1986; Maronna *et al*., 1979; Rousseeuw, 1984; Staudte & Sheather, 1990). Robust approaches have sound theoretical support and are a better remedy for the undesired effects of outliers than the various ad hoc rules for removing outliers that have been (and are currently) used in practice, see Kulinskaya *et al*. (1998), Bygrave & Benton (1998), Marazzi & Ruffieux (1999). Exploratory research (e.g., Kulinskaya *et al*.,1998; Marazzi *et al*., 1998) has shown the general advantages of robust methods. More specifically, research results from robust statistics have recently been adopted by the APDRG Association of Swiss hospitals (Marazzi & Ruffieux, 1999). The problems of heteroscedasticity of LOS have been addressed by recent research by Kulinskaya & Staudte (Kulinskaya *et al*., 1998; Kulinskaya *et al*., 2002; Kulinskaya et al., 2003), but other literature on this topic is sparse.

1.5.3 *Truncated maximum likelihood estimators.* The robust methods used in this study are known as truncated maximum likelihood (TML) methods. They can be successfully used for analysis of LOS and hospital costs and have been implemented in SPLUS routines, available from http://www.hospvd.ch/iumsp/home_a.htm. TML methods provide high efficiency and high breakdown point estimators for a class of regression models with asymmetric (or symmetric) error distribution. The details are available in Marazzi & Yohai (2004) and may be summarised as follows. Errors are assumed to belong to a location-scale family of distributions, such as the lognormal. A TML estimate is then computed in three steps. In the first step, a highly robust estimate is computed. In the second step, observations that are unlikely under the estimated model are rejected. In the third step, the maximum likelihood estimate is computed with the retained observations. For the second step, there are two methods of defining the rejection rule: a fixed and an adaptive one. In the fixed method, the standardised residuals outside a given quantile of their theoretical distribution (say, 0.99) are rejected. In the adaptive method, the cutoff is defined on the basis of the empirical log-likelihood distribution. This rejection rule results in the adaptively TML estimator or ATML-estimator. The ATML estimator achieves the best of both worlds: it is fully efficient at the model, and it also attains the maximum 50% break down point. We use both TML and ATML for the robust analysis of the LOS data in this paper.

2. Data sets and methods of analysis

2.1 *Data*

The data used were raw data from HRG Chapter A, the nervous system, comprising 34 HRGs (with two same-day HRGs, A07 and A08 excluded). The 1997/98 data set has 307227 cases, and the 1998/99 data set has 322135 cases. The spell duration, which is the basis for the HRG-based reference costs (DOH, 2004b), is used and referred to as the LOS.

All same-day cases (LOS $= 0$) were omitted, and only finished episodes for 'ordinary' patients with no missing values for any of the factors of interest, i.e. sex, age, provider, speciality, admission method, discharge destination or region were included. See HES Data Dictionary
http://www.publications.doh.gov.uk/hes/dictionary/index.html for definitions. Admission categories maternity and baby were also excluded, as they are very scarce for Chapter A.

2.2 *Trimming, transformations and data sets*

All analyses are performed using ln (LOS) as the dependent variable. However, before performing any analyses, the data could be trimmed to remove outliers. The standard NHS trimming process is performed on the raw data even though the main analysis is log transformed data. The data produced by this method is denoted data set 1. In this study, we also analysed a data set, data set 2, where the trimming was performed on log-transformed data. The trimming criterion for either method was 1.5 interquartile ranges from the upper quartile, with a separate trim point calculated for each HRG (National Casemix Office, 1997). The untrimmed data is denoted data set 0, and was used for all robust analyses. For 1997/98, data set 0 comprises 100% of the data; data set 1 has 198414 cases, 91.42% of the data; and data set 2 has 215848 cases, 99.45% of the data. Note that data set 2 includes 8% more cases than data set 1.

2.3 *Explanatory variables*

All data were adjusted for the variables, age, sex and HRG. The following additional factors were also investigated: admission method with three levels (emergency, elective, transfer); speciality with two levels (acute, geriatric); discharge destination with four levels (home, transfer NHS or Local Authority, including NHS and Local Authority run hospitals and residential accommodation; transfer non-NHS, including non-NHS non-LA hospitals and residential accommodation, or hospice; death); provider type with 8 levels used for clustering providers in NHSE (1999) (1 ‘acute specialist’ 2 ‘acute teaching’ 3 ‘large acute’ 4 ‘multiservice’ 5 ‘priority single service’ 6 ‘small/medium acute’ 7 ‘specialist community’ 8 ‘very large acute’); and region as of 1998 with 8 levels (1 Northern & Yorkshire, 2 Trent, 3 Anglia & Oxford, 4 North Thames, 5 South Thames, 6 South West, 7 West Midlands, 8 North West). Discharge destination is treated as an explanatory factor, even though it is obviously not known until after the event. This is to provide evidence that will enable LOS to be used more effectively for future costing by including discharge destination information.

2.4 *Methods of analysis*

There were two methods of analysis, standard (GLM) and robust (TML). The standard method was general linear model (GLM) applied to log-transformed data; it was performed on data sets 1 and 2. The robust method was TML regression assuming log-normal distribution (Marazzi & Yohai, 2004) and came in three versions according to the truncation criterion (90%, 95% or adaptive, ATML).

The following procedures were used for all data sets and methods of analysis.

2.4.1 *HRG, age and sex adjustment.* Current computation of the LOS indicator uses standardisation that weights the LOS data accordingly to HRG/age/sex distribution within a trust. If these weights are used and the data are then log-transformed, this is equivalent to an additive adjustment on the log scale. We did a similar adjustment by fitting a linear model to the log-LOS data, with sex as a factor and age as a covariate within each HRG, using either a standard or TML robust regression. The age effect was highly significant for every HRG, the sex effect only for some HRGs. Nevertheless, the sex effect was kept in all the models. We saved the residuals of this initial analysis and used them as an input data in the course of further analysis. For simplicity, we call these input data ‘the adjusted log-LOS’. If the current DOH methodology were used, the NHS Trust level means of these data (clustered by provider type) would be used to assess performance.

2.4.2 *Effects of admission, discharge, provider type, region and speciality.* To explore the effects of region, provider type, speciality, admission method and discharge destination on the adjusted log-LOS, we fitted a number of linear models including main effects and various two-way interactions.

We decided not to explore any higher level interactions for the sake of simplicity. All models were fitted using both standard and robust methods.

2.4.3 *Banding of the NHS Trusts.* A percentile method was used in NHSE (2002) to subdivide the performance of NHS Trusts on non-clinical indicators, including the LOS, into five bands. In this paper, Band 1 always signifies the best performance (the shortest adjusted LOS) and Band 5 the worst. The method splits the performance into five bands, split by the 10th, the 30th, the 70th and the 90th percentiles of the adjusted performance indicator, in our case the LOS. The rationale here is that such bands should be more robust than the simple ranking used in previous years. We repeat this procedure for the residuals from each of the fitted models, and compare the resulting bandings of NHS Trusts.

2.4.4 *Validation using the 1998/99 data.* TML and ATML robust regressions were used on the 1998/99 data set 0 to obtain 'new' parameter estimates, residual diagnostics and the R^2 values for all and for retained observations. Then, 'old' parameter values obtained from a model fitted for 1997/98 data were applied to the 1998/99 data set to obtain the residuals and the R^2 values. When discussing a fit of an 'old' model for the retained observations, observations are those retained in the 'new' model. The 'old' and the 'new' models are compared on their fit to the 1998/99 data.

3. Results

The fit of both standard and robust models to the 1997/98 data is considered first. Then the main effects of the explanatory factors are considered, as it turns out that including interactions provide little extra information on any model. Next, the implications of different methods for NHS banding are considered. Finally, models for 1997/98 data are validated against the 1998/99 data.

3.1 *Fit of models for 1997/98 data*

3.1.1 *Standard GLM.* Using the standard GLM, all main effects and all two-way interactions were significant, probably due to the data size. Analysis of residuals was originally performed on the log scale, but then the residuals were transformed back from the log scale to become ratios of observed to expected values of LOS. Table 1 provides total R^2 and several percentile points for the ratios of observed to fitted values. Higher R^2 and ratios closer to 1 indicate a better fit. The models are identified as follows. Step 1 has just age, sex and HRG. M0 is the model with step 1 plus the five main effects only. The further models are referred to by the total number of interactions, added in the following order: 1, admission by discharge; 2, admission by speciality; 3, admission by region; 4, region by provider type. The median of the ratios was 1 in all models.

From Table 1, it seems that the model with main effects fits the data better than just the step 1 HRG/age/sex adjustment, but the addition of interaction terms improves the fit of the models very little. It was also predicted that any model should fit better for data set 1 than data set 2, as data set 1 has more outliers removed. This prediction was born out in terms of percent variance accounted for. For M0, R^2 was 8.4% for data set 1, but 6.6% for data set 2. The percentiles of the residuals are also instructive. For data set 1, the central 50% of the ratios of observed divided by expected values are within a two-fold range of error, and the central 80% is within a four-fold range. For data set 2, the central 80% is much tighter within a three-fold range. It is also noticeable that the maximum residual is of the order of 10 times as large for data set 1 as for data set 2.

TABLE 1 *Fit of standard analysis for data sets* 1 *and* 2 *for models M*0*, M*1*, M*3*, M*4*, M*10

			Percentiles for ratios of observed/expected estimates							
Model	Data set	R^2	Min	5	10	25	75	90	95	Max
Step1	1		0.01	0.17	0.27	0.48	2.04	4.01	5.96	149.90
M0	1	0.084	0.02	0.19	0.28	0.50	1.98	3.75	5.50	219.20
M1	1	0.087	0.02	0.19	0.28	0.50	1.98	3.74	5.47	157.59
M3	1	0.090	0.02	0.19	0.28	0.50	1.98	3.72	5.44	154.47
M4	1	0.093	0.01	0.19	0.28	0.50	1.98	3.71	5.42	157.59
M10	1	0.096	0.01	0.19	0.28	0.50	1.98	3.71	5.41	151.41
Step1	2		0.06	0.20	0.31	0.52	1.99	3.29	4.16	24.53
M0	2	0.066	0.03	0.21	0.32	0.54	1.93	3.14	4.02	23.57
M1	2	0.068	0.02	0.21	0.32	0.54	1.92	3.14	4.02	23.57
M3	2	0.070	0.02	0.22	0.32	0.54	1.92	3.14	4.01	23.81
M4	2	0.072	0.02	0.21	0.32	0.54	1.92	3.13	4.01	23.10
M10	2	0.075	0.02	0.22	0.32	0.54	1.92	3.12	4.00	24.53

TABLE 2 *Fit of robust TML analysis for data set* 0 *for models M*0 *and M*4

		Retained		Percentiles for ratios of observed/expected estimates							
Cutoff	Model	% of observation	R^2	Min	5	10	25	75	90	95	Max
1.65	M0	88.4	0.137	0.17	0.27	0.35	0.55	1.82	2.95	3.72	5.70
1.65	M4	88.4	0.153	0.16	0.27	0.36	0.56	1.81	2.93	3.68	5.87
1.96	M0	93.4	0.113	0.12	0.23	0.32	0.53	1.89	3.26	4.32	8.58
1.96	M4	93.3	0.126	0.11	0.24	0.32	0.54	1.88	3.24	4.29	8.17
Adaptive	M0	98.3	0.093	0.06	0.19	0.29	0.51	1.96	3.66	5.26	17.46
Adaptive	M4	98.2	0.105	0.05	0.20	0.29	0.51	1.96	3.62	5.17	16.95

Visual inspection of the box plots for residuals by HRG showed that the distributions are symmetric with many outliers and with visibly different interquartile range (IQR) between HRGs. This supports the need for robust methods.

3.1.2 *Robust TML.* Only two models were fitted, the main effects model, M0, and the M4 model with four interactions. Analyses were performed with a 90% criterion, a 95% criterion and ATML with an adaptive criterion. Analysis of residuals (transformed into ratios of observed to expected values) is shown in Table 2. As with the standard analysis, there is little advantage in including interaction terms. The R^2 increase for M4 over M0 is less than 1.6% for the 90% cutoff of 1.65, and effectively zero for the 95% cutoff of 1.96 or for the adaptive cutoff. As would be expected, decreasing the cutoff value results in the better quality of fit, as measured both by the R^2 and the residuals. The 1.96 cutoff is little worse than the 1.65 cutoff for R^2 and the central 50% of residuals. For more extreme values, the higher advantage of the higher proportion of retained observations must be balanced against higher residuals for extreme values.

3.2 *Comparison of robust TML and standard GLM fit*

Only the fit for model M0 is considered, as this is the model that we recommend for this data. However, results for other models would be very similar. It is sensible to compare TML with 1.96 cutoff with

TABLE 3 *Comparison of fit of robust TML and standard GLM for model M*0

	Retained		Percentiles for ratios of observed/expected estimates							
Model	% of observation	R^2	Min	5	10	25	75	90	95	Max
GLM data set 1	91.4	0.084	0.02	0.19	0.28	0.50	1.98	3.75	5.50	219.20
TML, 1.96	93.4	0.113	0.12	0.23	0.32	0.53	1.89	3.26	4.32	8.58
GLM data set 2	99.5	0.066	0.03	0.21	0.32	0.54	1.93	3.14	4.02	23.57
ATML	98.3	0.093	0.06	0.19	0.29	0.51	1.96	3.66	5.26	17.46

TABLE 4 *Ratio of days at worst level to days at best level for five factors according to three methods of analysis*

	Method of Analysis		
Factor	TML, 1.96	ATML	GLM
Discharge destination	3.66	3.85	3.72
Discharge (excluding death)	2.36	2.27	2.18
Admission method	1.47	1.50	1.49
Provider type	1.40	1.39	1.36
Region	1.22	1.24	1.22
Speciality	1.20	1.18	1.40

data set 1 and ATML with data set 2, as this compares models with a similar proportion of retained observations. Table 3 summarizes these comparisons. In terms of percentage of variance accounted for the robust methods have an R^2 advantage of nearly 3%. The central 50% of the ratios of observed to expected values are similar and around twofold for all models. The striking advantage of the robust TML methods is apparent in the more extreme residuals, especially for the maximum, as may be seen in Table 3.

3.3 *Effects of explanatory factors*

The analysis of fit for both robust and standard models shows that interactions, although significant, have very small effect sizes. Consequently, only the results for model M0 with no interactions are considered. Three different methods of analysis were used: TML with 1.96 cutoff, ATML and standard GLM for data set 2. The original analyses were on ln (LOS), but the obtained parameters have been back transformed with an exponential transformation to generate parameters that represent the ratio of LOS days at a particular level to LOS days for the reference level. In order to gauge the overall effect of each factor, the ratio of number of days at the worst level of a factor to number of days at the best level of that factor were calculated and are shown in Table 4.

All methods show the strongest effect for discharge destination, even if discharge due to death (notorious for short LOS) is excluded, followed by admission method. The robust methods show the next strongest effect for provider type, while GLM shows the next strongest effect for speciality. Clearly, the method chosen matters.

Table 5 shows details of the parameters for all levels for each factor in terms of ratio of days at a level to days at the reference level, together with the width of their 95% confidence intervals (CIs). Although the parameters are broadly similar, there are some quite large differences between GLM and the robust models. In general, the largest parameters tend to be even further from 1 for GLM than for TML.

TABLE 5 *Parameters for the main effects of five factors according to three methods of analysis*

	Days/reference days			Width of CI (days/reference days)		
	TML, 1.96	ATML	GLM	TML, 1.96	ATML	GLM
Admission method						
Elective	0.85	0.85	0.95	0.02	0.02	0.03
Emergency	Reference					
Transfer	1.25	1.27	1.41	0.03	0.04	0.03
Discharge destination						
Dead	0.59	0.55	0.60	0.02	0.02	0.02
Transfer: NHS or LA	0.91	0.93	1.02	0.03	0.03	0.02
Home	Reference					
Transfer: non-NHS or LA	2.15	2.12	2.23	0.09	0.07	0.13
Region						
South West	0.93	0.92	1.03	0.03	0.01	0.04
Anglia & Oxford	0.94	0.94	1.05	0.03	0.03	0.04
Trent	0.95	0.94	1.04	0.03	0.03	0.04
Northern & Yorkshire	Reference					
South Thames	1.05	1.04	1.14	0.03	0.04	0.05
North West	1.07	1.07	1.18	0.03	0.01	0.04
West Midlands	1.08	1.08	1.18	0.03	0.04	0.03
North Thames	1.13	1.14	1.26	0.03	0.04	0.03
Provider type						
Large acute	0.91	0.92	1.03	0.02	0.03	0.03
Small/medium acute	0.94	0.94	1.07	0.03	0.04	0.06
Very large acute	0.95	0.95	1.07	0.02	0.03	0.05
Acute specialist	Reference					
Multi service	1.01	1.02	1.15	0.03	0.01	0.06
Acute teaching	1.03	1.03	1.15	0.02	0.01	0.05
Specialist community	1.28	1.28	1.40	0.01	0.06	0.09
Speciality						
Acute	Reference					
Geriatric	1.20	1.18	1.40	0.01	0.02	0.03

Although the effects are broadly similar for the robust and standard methods, it should be noted that the mean standard error of the estimates in the original logged metric was substantially larger for GLM (mean s.e. $= 0.011$) than for the TML (mean s.e. $= 0.006$).

The main effects may be summarised as follows. Discharge destination has a strong influence on LOS. Death has the shortest LOS, a familiar finding. Discharge to a private hospital or residential home has the longest LOS, more than twice the LOS for home discharge, which is quite similar to the LOS for discharge to public hospital or residential home. Admission method is also influential. Unsurprisingly, elective patients have lower LOS than acute patients, who in turn have lower LOS than transfer patients. Provider type also matters, especially according to the TML analyses. The biggest effect is the longer LOS for specialist community hospitals. The geriatric speciality has longer LOS than acute, even after adjustment for HRG and age. It may be that more chronic cases are sent to geriatric. Finally, the effects of region are small. Since region changes every 10 minutes, and is in any event a poor proxy for poverty, it is probably not worth including in future models.

3.4 *Banding of the NHS Trusts*

After adjusting the LOS in each of the fitted models, we ranked the 266 NHS Trusts, and compared the resulting rankings. The agreement was high but not perfect, with Spearman correlations varying from 0.900 to 0.985.

Then we subdivided the NHS Trusts into 5 bands, following the NHSE (2002) methodology. Agreement of these bandings was assessed by Cohen's (1968) kappa. The agreement between step 1 (age, sex and HRG) and the M0 models, however analysed, was low: 0.26 for GLM; 0.32 for TML, 1.96; and 0.33 for ATML. Consequently basing bandings on percentiles that ignore key factors such as admission and discharge type are likely to be unreliable and highly misleading. The agreement for M0 between GLM and the robust method was high, but not high enough: kappa (GLM, TML, 1.96) $=$ 0.73, kappa (GLM, ATML) $=$ 0.86. Kappa between M0 and the interaction models, M4 and M10, were also low. So altogether bandings according to percentiles do not make a lot of sense.

3.4.1 *Comparison of LOS and survival rates.* Since LOS is so much lower for death than other discharge types we decided to explore the relationship of LOS and survival rates in more detail. The survival rate is obviously another important hospital performance indicator, and is known to be related to LOS. Indeed the Dr Foster Hospital Guide (2004) uses survival rates that are adjusted by the LOS to be comparable between providers.

Data was aggregated at the NHS Trust level. Then the average adjusted LOS for survivors was compared with the overall average adjusted LOS, and found to be higher in 194 NHS Trusts and lower in only 51 NHS Trusts. This confirms our conclusion that the LOS is on average shorter for the deceased.

Then we calculated Spearman correlation coefficient ρ between the average casemix-adjusted LOS (just sex, age and HRG) and the survival rate ($\rho = -0.175$, $p = 0.004$). This correlation of the adjusted LOS to the survival rate disappeared after adjustment by all the models (both parametric and robust, regardless of the cutoff). However, a more in-depth analysis showed that survival rates tend to be lower in adjusted LOS percentile bands 2–4, whereas both band 1 and band 5 have higher median survival rates. This applies to adjustment of the LOS by all fitted models with subsequent subdivision of the NHS Trusts into 5 bands, see Table 6.

Thus our analysis demonstrated a new pattern in the relationship between the LOS and survival rates. In a number of NHS Trusts a high average LOS is explained by a better survival of (more complicated?)

TABLE 6 *Median and quartile survival rates by bands defined by adjusted LOS*

Model	Casemix-adjusted LOS			M0, GLM			M0, TML, 1.645			M10, GLM		
Band	Median	25%	75%	Median	25%	75%	Median	25%	75%	Median	25%	75%
1	0.9099	0.8698	0.9948	0.8951	0.8429	0.9948	0.8743	0.8429	0.9948	0.9085	0.8712	1.000
2	0.8822	0.8536	0.9107	0.8714	0.8446	0.9109	0.8680	0.8253	0.9006	0.8702	0.8456	0.9098
3	0.8712	0.8435	0.8967	0.8682	0.8407	0.8962	0.8682	0.8408	0.8993	0.8710	0.8365	0.9025
4	0.8646	0.8149	0.9044	0.8651	0.8181	0.8986	0.8825	0.8428	0.9103	0.8624	0.8364	0.8849
5	0.8359	0.8123	0.9472	0.8875	0.8423	0.9472	0.8858	0.8336	0.9396	0.8852	0.8176	0.9567

TABLE 7 *Comparison of the 'old' and 'new' main effects models on the* 1998/99 *data*

				Percentiles for ratios of observed/ expected estimates							
Analysis	Retained % of observation	Parameters	R^2	Min	5	10	25	75	90	95	Max
TML, 1.96	93.4	New 98/99	0.117	0.12	0.23	0.32	0.53	1.88	3.28	4.38	8.50
		Old 97/98	0.116	0.12	0.23	0.32	0.54	1.90	3.30	4.41	8.58
ATML	98.3	New 98/99	0.097	0.06	0.20	0.28	0.51	1.96	3.68	5.29	18.36
		Old 97/98	0.096	0.06	0.19	0.28	0.51	1.96	3.68	5.28	18.54
	99.5	New 98/99	0.086	0.02	0.18	0.28	0.50	1.98	3.78	5.53	188.67
		Old 97/98	0.081	0.02	0.18	0.28	0.52	2.01	3.84	5.65	83.93

cases. This pattern is masked by a seemingly straightforward negative correlation between the casemix-adjusted LOS and survival when no additional factors are taken into account.

3.5 *Validation on 1998/99 data*

The fit of the models based directly on the 1998/99 data is compared to that of the models based on the parameters obtained from analysis of the 1997/98 data. The results are shown in Table 7. The results are highly consistent across years for all models, as demonstrated by nearly identical values for all fit measures. This means that the effects of the 5 factors of interest are genuine and remain comparatively constant over time. We do not reproduce the parameter estimates for the 1998/99 data, but they are also very close to those given in Table 4 for the 1997/98 data. The implication is that the factors investigated, especially those of admission type and discharge destination need to be taken into account in any future use of LOS for either performance assessment, or casemix funding.

4. Discussion

We first discuss the implications of our substantive findings on the five factors that are not currently used to adjust LOS, when used either for costing or as a high-level performance indicator. Then we briefly discuss the methodological findings with respect to robust and standard methods of analysis.

4.1 *Implications for the use of LOS for costing and as a performance indicator*

The results of our analysis showed a number of rather well known effects of various factors on the LOS. The main problem with the whole area of health performance measurement and casemix funding is that the knowledge possessed by many health managers is continuously ignored when devising new performance measures or even a new system of public health funding.

Since the new casemix based funding is being introduced in the UK from 2005/06 we looked at the published cost-weights DOH (2004b) to see how various factors were implemented in the new funding. We then compared these cost-weights with what would be recommended based on our findings. Admission method has two different cost-weights for each HRG: one for acute and one for elective admissions. There is no separate cost-weight for a transfer, though our analysis shows these will be the cases with the longest LOS. Discharge destination is completely ignored in the new system. It is arguable that the infrastructure in the area will define how long the transfers stay in hospital, and the local Health Authority should suffer those expenses. But what is really very important is that the death usually happens early, so the LOS is shorter for deceased patients. Since the whole casemix funding system encourages short LOS, trusts with high mortality may be doing the best under the new system. This is certainly not desirable from the patients' point of view. Our results for provider type are contrary to the assumption of previous performance measurement exercises, and support the same payment for all. Special community hospitals are the exception with at least a 20% longer LOS than other providers. There is a considerable difference between the acute and geriatric specialties. Partly it is explained by the existing infrastructure, but the complexity may also be different. This factor requires further investigation. As already noted, including region in LOS analysis is dubious. Categorised patient postcodes are a better proxy for poverty.

4.2 *Comparing standard and robust methods of analysis*

The first point to note is that if one is using standard methods then log (LOS) based trimming is more efficient than raw LOS based trimming, as it keeps far more of the data without losing sensitivity. The robust TML models are attractive, particularly ATML, since they produce more accurate estimates as shown by the smaller standard errors for all estimated parameters. They also appear to have better fit characteristics in terms of variance accounted for, for equivalent proportion of retained cases.

5. Summary

Our analyses show that there are at least four factors that have a strong influence on LOS and all merit adjustment when using LOS for important decisions such as costing or performance measurement. Discharge destination and admission methods are particularly important, and not fully adjusted for even in the most recent NHS proposals.

Furthermore, robust TML methods have attractive advantages over standard GLM methods. For any given model, they produce a better fit in terms of variance accounted for. Furthermore, parameters are more accurately estimated, with standard errors reduced by approximately 50%.

Acknowledgements

We are grateful to Alfio Marazzi for his advice on robust methods and to Alex Randriamiharisoa for his help with the use of the TML programs. The authors thank the Department of Health, Directorate of Health and Social Care, Midlands and East of England, for support of this research, contract number HSR 1100/8; RCC 33039.

References

Briggs, A. & Gray, A. (2000) Using cost effectiveness information. *BMJ*, **316**, 246.

Bygrave, S. & Benton, P. (1998) To trim or not to trim. Patient Classification Systems/Europe (PCS/E). *Proceedings of the 14th International Working Conference*, 1–3 october 1998, Manchester, England, pp. 352–365.

Cohen, J. A. (1968) Weighted *kappa*: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, **70**, 213–220.

Deyo, R. A., Diehl, A. K. & Rosenthal, M. (1986) How many days of bed rest for acute low back pain? *N. Engl. J. Med.*, **315**, 1064.

DOH (2004b) NHS Reference Costs 2003 and National Tariff 2004 ('Payment by Results Core Tools 2004') (http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/Publications PolicyAndGuidanceArticle/fs/en?CONTENT_ID=4070195&chk=UzhHA3).

DR FOSTER HOSPITAL GUIDE (2004) Hospital guide (http://www.drfoster.co.uk/hospital_guide/main/methodology.asp).

FETTER, R. B. (ED.) (1991) *DRGs: Their Design and Development*. Ann Arbor, MI: Health Administration Press.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986) *Robust Statistics. The Approach Based on Influence Functions*. New York: John Wiley & Sons.

HUBER, P. J. (1981) *Robust Statistics.* New York: John Wiley & Sons.

KULINSKAYA, E., KNIGHT, E., KORNBROT, D. E. & BENTON, P. (2001) The use of log and power transformations in the analysis of length of stay data. *Casemix Quarterly*, **3**, 79–89.

KULINSKAYA, E., STAUDTE, R. G. & GAO, H. (2003) Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Commun. Stat. Theory Methods*, **32**, 2353–2371.

KULINSKAYA, E., STAUDTE, R. G., HALES, C. & LEUNG, C. (2002) Does weighted ANOVA and weighted R^2 work on real LOS data? *Patient Classification Systems Europe (PCS/E). 18th International Case Mix Conference Proceedings*, 2–5 October 2002, Innsbruck, Austria, PCS/E pp. 23–38.

KULINSKAYA, E., STAUDTE, R. G., ZHANG, X. M. (1998) A robust method for reviewing variance explained by diagnosis related groups. *Patient clasification Systems/Europe (PCS/E). 14th International Working Conference Proceeding Manual*, 3 October 1998, Manchester, UK, pp. PCS/E 185–190.

LAVE, J. R. & LEINHARD, S. (1976) The cost and length of a hospital stay. *Inquiry*, **13**, 327–343.

MARAZZI, A., PACCAUD, F., RUFFIEUX, C. & BEGUIN, C. (1998) Fitting the distributions of length of stay by parametric models. *Med. Care*, **36**, 915–927.

MARAZZI, A. & RUFFIEUX, A. (1999) The truncated mean estimate of the mean of an asymmetric distribution. *Comput. Stat. Data Anal.*, **32**, 79–100.

MARAZZI, A. & YOHAI, V. (2004) Adaptively truncated maximum likelihood regression with asymmetric errors. *J. Stat. Plann. Inference*, **122**, 271–291.

MARONNA, R. A., BUSTOS, O. H. & YOHAI, V. J. (1979) Bias- and efficiency-robustness of general M-estimators for regression with random carriers. *Smoothing Techniques for Curve Estimation* (T. Gasser & M. Rosenblatt eds). *Lecture Notes in Mathematics 757*. Berlin: Springer, pp. 91–116.

MUSCHLIN, A. I., BLACK, E. R. & CONNOLLY, C. A. (1991) The necessary length of stay for chronic pulmonary disease. *JAMA*, **266**, 80–83.

National Casemix Office (1997) *Version 3 Healthcare Resource Group Documentation Set*, vol. 2, Compendium. Winchester: National Casemix Office.

NHSE (1999) *Quality and Performance in the NHS: High Level Performance Indicators.* Finance and Performance Assessment, NHS Executive, June 1999.

NHSE (2000) *NHS Performance Indicators*. Finance and Performance Assessment, NHS Executive, NHS Cat. No. 21946, July 2000.

NHSE (2002) *NHS Performance Indicators*. February 2002 (http://www.performance.doh.gov.uk/hsperformanceindicators/2002/index.html).

ROE, CH. J., KULINSKAYA, E., BRISBANE, M., BROWN, R. & BARTER, C. (1996) A methodology for measuring clinical outcomes in an acute care teaching hospital. *J. Qual. Clin. Pract.*, **16**, 203–214.

ROE, C. J., KULINSKAYA, E., DODICH, N. & ADAM, W. R. (1998) Comorbities and prediction of length of hospital stay. *Aust. N. Z. J. Med.*, **28**, 811–815.

STAUDTE, R. G. & SHEATHER, S. J. (1990) *Robust Estimation and Testing*. New York: John Wiley & Sons.