Health Care Management Science 6, 27–35, 2003

Relevance of Outlier Cases in Case Mix Systems and Evaluation of Trimming Methods

FRANCESC COTS *, DAVID ELVIRA and XAVIER CASTELLS
Municipal Institute of Health, Health Service Research Unit, Hospital del Mar, Passeig Marítim 25-29, E-08003 Barcelona, Spain
E-mail: fcots@imas.imim.es

MARC SÁEZ
University of Girona, Department of Economics, Campus de Montilivi, 17071 Girona, Spain

Abstract. *Objectives*: To determine the most appropriate outlier trimming method when the main source of information for case mix classification is length of stay (LOS) because cost information is unavailable. *Methods*: Discharges (35,262) from two public hospitals were analysed. LOS and cost outliers were calculated using different trimming methods. The agreement between cost and LOS trimming was analysed. *Results*: The trimming method using the geometric mean with two standard deviations (GM2) showed the highest level of agreement between cost and LOS and revealed the greatest proportion of extreme costs. Nearly 5% of cases were outliers, containing 16% of total LOS. This was the best approximation to 18% of extreme cost because when GM2 was applied to LOS, 88% of outlier cost was revealed. *Conclusions*: The methods were analysed because they are the most frequently used but the same methodology could be employed to compare other outlier determination methods. Outliers should be calculated because they ought to be valued differently from inlier cases.

Keywords: outliers, trimming methods, hospital cost, cost analysis, case mix

1. Introduction

Hospital cost analysis has advanced despite the difficulties of delimiting the product resulting from hospital activity. The total amount of activity, intermediate products and ways of approaching different diseases are interrelated in each hospital patient. Consequently, the associated cost is extremely difficult to evaluate.

At the beginning of the 1980s, systems to define hospital product began to be used that attempted to contain all activities, products and diagnoses in a limited number of groups. The central axis of these systems is the inpatient who is assigned to a product group related to diagnosis and to other criteria concerning severity of illness, complications, comorbidities, or age. To determine the relationship between total hospital cost and total hospital product, a simple linear function is used [4,5]:

$$C_t = \sum_{i=1}^{n} k_i \cdot Y_i, \tag{1}$$

where k_i is the number of cases and Y_i is the value assigned to the product i.

The main advantage of these patient classification systems is that they eliminate the imprecision inherent in the multiproduct nature of hospitals where collective action in any given patient results in infinite possibilities [18]. In practical terms, this means that these systems are able to establish a dialogue between the provider and the purchaser. The establishment of a purchase–provision relationship for a large number of patients means that inter-patient cost variation can be approached through patient classification systems and that a price can be assigned to each category [27]. Much of the unexplained variability that arises when the care a patient receives is more expensive than the invoice is compensated for by the fact that in another patient the opposite occurs. Similarly, if a whole diagnosis related group (DRG) is underfinanced for a given hospital, this is compensated for by the over-financing of another DRG.

The main limitation of these systems is that they assume that the costs associated with a particular are usually the same and therefore, that the activities associated with patients in a particular group are also the same.

Several analyses have found these systems' ability to explain cost variation to be limited [6,32,34,36]. The reduction of the cost of one group of patients to the mean cost of patients belonging to a particular group is a very significant reduction in the complex framework of hospital product [25].

Little more than half of DRGs, the most commonly used patient classification system, represents more than 90% of the activity of a general hospital [14]. Not more than 250 mean values should explain the cost variation related to 20,000 patients in a particular hospital or to 2,000,000 discharges from a public hospital system.

Another clear limitation of patient classification systems lies in the fragmentation of patient care. These systems include the field of inpatient hospital care, omitting other hospital activities (outpatient, emergency, and day-hospital care) because of the added complexity of not having the clear temporal demarcation of patient admission and discharge.

* Corresponding author.

1.1. Lognormal cost distribution per hospital patient

When determining the mean cost associated with each DRG in order to take a decision about cost, the mean value may not be the most appropriate tool with which to value all the patients in a particular group.

From a mathematical point of view, cost distribution per patient has a minimum value of 0 and a maximum value tending towards the infinite. Patient cost distribution is heavily skewed to the right, signifying a form of cost distribution in which the mean value is subject to tensions that could lead to overvaluation. Consequently, as several authors have pointed out, cost function distribution is lognormal [3,24].

1.2. Financial risk

Patient classification systems have been extensively developed due to their use as an instrument of hospital payment. Obviously, the application of mean values to an entire national health system supposes a financial risk for hospitals whose case mix does not fit that of the whole system [16,19,20,26]. Thus, if the number of patients with a higher cost than the value recognised by the corresponding DRG exceeds that of patients with a lower cost than the standard value, the result will be overall undervaluation of the hospital product.

1.3. Outlier cases

DRG groups incorporate patients far removed from the resource use of most patients belonging to the same group [1]. The consequent valuation according to the mean of the patients in the group incorporates the tendency of these outlier cases. This effect, known as masking, leads to the overvaluation of the mean value of this category. On the other hand, the existence of these cases in resource use (length of hospital stay (LOS) or cost) means that, when standard external valuations are used to identify them, their cost is undervalued.

The Medicare system in the United States [7,8,21], the National Health Service in the United Kingdom [30,31,33] and DRG-based clinical analysis systems [9,10,28] eliminate these values to determine standard mean values, which makes robust inter-hospital and inter-year comparisons possible. Medicare uses a differential payment for these cases.

1.4. Determination of outlier cases

To detect outlier cases, several trimming methods can be used that yield different results [22,23]. Of these methods, two stand out. One group is based on the distribution of the elements that compose the group to be analysed. These methods attempt to make the arithmetic mean more robust. They use a multiple of the standard deviation of normal distribution to designate the trimpoint of outlier cases:

$$\text{trimpoint} = \text{mean} + a \cdot \text{standard deviation}, \qquad (2)$$

where a is the parameter that multiplies standard deviation.

A second group of trimming methods is based in the inter-quartile range so that a multiple of the range between the 25th and 75th percentiles is added to the 75th percentile:

$$\text{trimpoint} = 75\text{th percentile} + a \cdot \text{inter-quartile range}. \qquad (3)$$

In addition to the difference in conception, parametric or non-parametric, these two methods also differ in the parameter used. The parameter is the number of times that either the inter-quartile range is added to the 75th percentile or the standard deviation is added to the mean. If parametric methods are used, the cost per patient distribution for each category must be normal, otherwise the parameters used will be skewed. Because cost function distribution is lognormal, a logarithmic transformation can be performed and subsequently applied to the parametric method used. In this case, once the trimpoint has been found, the transformation obtained reverts to its original value.

In practice, the geometric mean and the standard deviation of the original distribution are used. The geometric mean equals the arithmetic mean calculated over the logarithmic transformation. Thus, the parametric methods used by Medicare until 1997 are given by:

$$\text{trimpoint} = \text{geometric mean} + a \cdot \text{standard deviation}. \qquad (4)$$

The aim of this study was to determine the most satisfactory outlier trimming method and to analyse its relevance in the distribution of financial risk among hospitals. The methodology was based on the hypothesis that the most satisfactory trimming method is that which shows the greatest agreement when applied to LOS and to costs. This is a practical hypothesis because in most European countries resource use is assimilated by LOS and only rarely is systematic information on cost per patient available in hospital information systems. Per patient cost information is the main outcome measure of the hospital process. When this information is unavailable, LOS is used. LOS is a physical measure of hospital resource use and incorporates most intermediate products used in the treatment of the inpatient.

This study makes use of per discharge cost information from two public hospitals in Barcelona from 1995 to 1996. In Spain, no other hospitals have per-patient cost information at their disposal, a situation that is also true of most hospitals in European health systems.

2. Material and methods

The discharges of patients admitted to the two teaching hospitals during a two-year period between 1995 and 1996 were analyzed. These hospitals showed a relative case mix index (calculated using Medicare's DRG-weights) of 1.14 with respect to the 600,000 discharges from the Catalan health system in 1996. These two hospitals have the eighth highest case mix index in the Catalan public hospital network, which contains 70 hospitals.

2.1. Determination of cost per patient

The Municipal Institute of Health uses a hospital cost accounting system based on full costing allocation [11,35]. This system ensures that the hospitals' total costs are distributed among the patients. Allocation is based on directly assigning the cost of the following services to the patient: Laboratory, Pharmacy, Radiology, Nuclear Medicine, Pathological Anatomy, and Prosthesis [13]. The information systems contain exhaustive data on human resources and their activity: storage, planning of admission, ambulatory and emergency care, operating rooms, diagnostic and complementary tests, and inter-hospital consultations. This information creates and automatically updates the cost drivers for overheads [37,38]. Teaching activity is assessed according to the agreements between the health institutions and the universities. Research activity is valued on the basis of the impact factor of the clinical staff's publications.

2.2. Calculation of trimming methods

A database with costs and LOS for 35,262 patients was constructed. For each DRG the trimpoints for the following equations were calculated both for costs and for LOS:

$$\text{GM} + 2 \cdot \text{SD (referred to hereafter as GM2)}, \quad (5)$$

$$\text{GM} + 3 \cdot \text{SD (referred to hereafter as GM3)}, \quad (6)$$

$$75\text{th percentile} + 1.5 \cdot \text{IR (referred to hereafter as IQ15)}, \quad (7)$$

$$75\text{th percentile} + 2 \cdot \text{IR (referred to hereafter as IQ20)}, \quad (8)$$

where GM is the geometric mean, SD is the standard deviation and IR is the inter-quartile range.

2.3. Behaviour of the different trimming methods

The main difference between parametric and non-parametric methods is the influence of intra-DRG variability on the value of the trimpoint. The behaviour of each trimming method in relation to the degree of intra-DRG variability was analysed.

The total amount of cost and LOS that remained above the trimpoint was valued to determine the potential impact of an extra payment recognising outlier cost on the total cost of the public hospital network.

2.4. Analysis of agreement

Contingency tables showing four possibilities were constructed: one representing agreement between outlier cases in terms of cost and LOS, one representing agreement between inlier cases and two representing non-agreement (table 1).

Several "case by case" agreement tests for each method used were applied to these contingency tables [1,29]. These tests measure the proportion of agreement and the level of bias in the cells showing non-agreement. We considered cost as the observed variable and LOS as its estimation.

Table 1
2×2 contingency table.

Costs	Length of stay		
	Inliers	Outliers	Total
Inliers	inliers (a)	false positives (b)	$(a) + (b) = c1$
Outliers	false negatives (c)	outliers (d)	$(c) + (d) = c2$
Total	$(a) + (c) = e1$	$(b) + (d) = e2$	$(a) + (b) + (c) + (d) = T$

2.5. Kappa's coefficient

This chance-corrected index was used to determine the degree of agreement between the results for LOS and for costs:

$$\kappa = (\text{PO} - \text{PC})/(1 - \text{PC}), \quad (9)$$

where PO is proportion of observed agreement (in table 1: $\text{PO} = (a + d)/T$) and PC is proportion of chance agreement (in table 1: $\text{PC} = (e1 \cdot c1 + e2 \cdot c2)/T^2$).

The Kappa coefficient oscillates between negative values and 1. If the agreement between observation methods A and B is equal to what could be expected by chance, then $\kappa = 0$. When $\text{PO} = 1$ (and consequently there is perfect agreement) then $\kappa = 1$. Landis and Koch suggest an interpretation of this coefficient, classifying agreement into: bad (<0), poor (0–0.20), average (0.21–0.40), moderate (0.41–0.60), substantial (0.60–0.80) and almost perfect (0.81–1). A limitation of the Kappa coefficient is that unequal symmetrical distributions yield low Kappa coefficients despite higher degrees of agreement.

2.6. Sensitivity index

The sensitivity index is the proportion of LOS outlier cases detected that agreed with cost outlier cases (standard). The result is expressed in values between 0 and 1:

$$\text{sensitivity} = d/c2. \quad (10)$$

2.7. Specificity index

The specificity index is the proportion of LOS inlier (not outlier) cases that agreed with cost inlier cases (standard). The result is expressed between 0 and 1:

$$\text{specificity} = d/c1. \quad (11)$$

2.8. Youden's index

Youden's index includes the sensitivity and specificity indexes. It is given by equation (12) [17]. The result is expressed in values between 0 and 1:

$$Y = S + E - 1. \quad (12)$$

The sensitivity and specificity indexes and Youden's index do not correct for the effect of chance. However, the asymme-

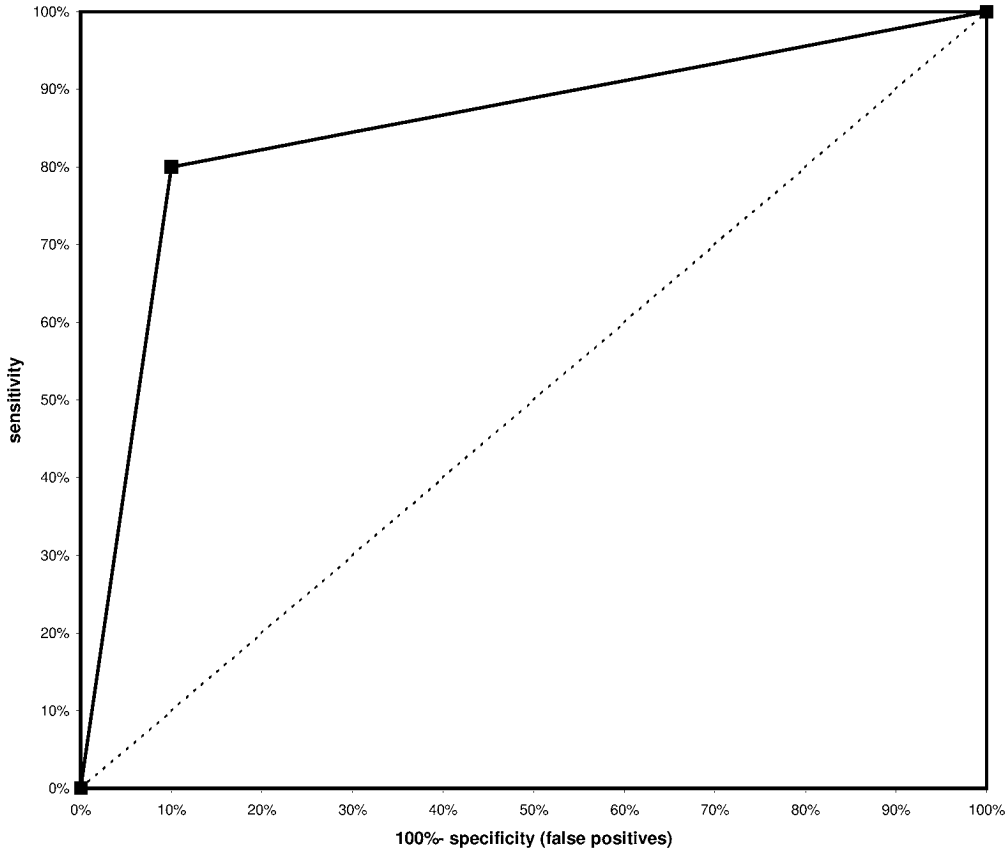

Figure 1. The receiver operating characteristic (ROC) curve is the relationship established between sensitivity and the complementary specificity. The greater area below the ROC curve, the greater agreement due to higher sensitivity and lower lack of specificity.

try of the contingency table affects the value of the sensitivity and specificity indexes less than that of Kappa's coefficient.

2.9. Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve is the relationship established between sensitivity and the complementary specificity [1]. Figure 1 shows that the greater the area below the ROC curve, the greater the agreement on adding higher sensitivity and lower lack of specificity. The area below the ROC curve can be calculated by applying equation (13). The maximum value which can be obtained by this index is 100, equivalent to leaving an area of 100% below the curve:

$$\mathrm{ROC} = 1 - U/(c1 \cdot c2), \tag{13}$$

where U is the value of the Mann–Whitney non-parametric test for the outlier–inlier LOS variable grouped according to the outlier–inlier cost variable and $c1$ and $c2$ are the number of cost inlier cases and cost outlier cases, respectively.

The Mann–Whitney non-parametric test relates the explanatory variable with the grouping of the independent variable. The independent variable is whether or not the case is of extremely high cost and the explanatory variable is whether or not the case is of extremely high LOS.

2.10. McNemar's test

McNemar's test determines whether there is systematic bias in the cases showing non-agreement, as shown in cells b (false positives) and c (false negatives) in table 1. Systematic bias exists when one type of error predominates. McNemar's test statistic is given by:

$$\chi^2 = (b - c)/(b + c). \tag{14}$$

The critical point of this statistic is $\chi^2_{1,\alpha}$. Above this level, systematic bias at a significance level of α is believed to exist.

2.11. Cost associated with outlier cases determined by LOS

The quality of the trimming method should be analysed according to the volume of extreme costs identified by the method when applied to the variable of LOS. Independently of the agreement in number of cases, the volume of costs associated with these cases is the most relevant factor to consider when choosing a trimming method. The quality indicator created enabled calculation of the percentage of outlier costs identified when the trimming method was applied to the variable of LOS.

95% confidence intervals were calculated for all the indexes used in the analyses. The bootstrapping method with 1000 repetitions was used to calculate the confidence intervals for all non-parametric measures.

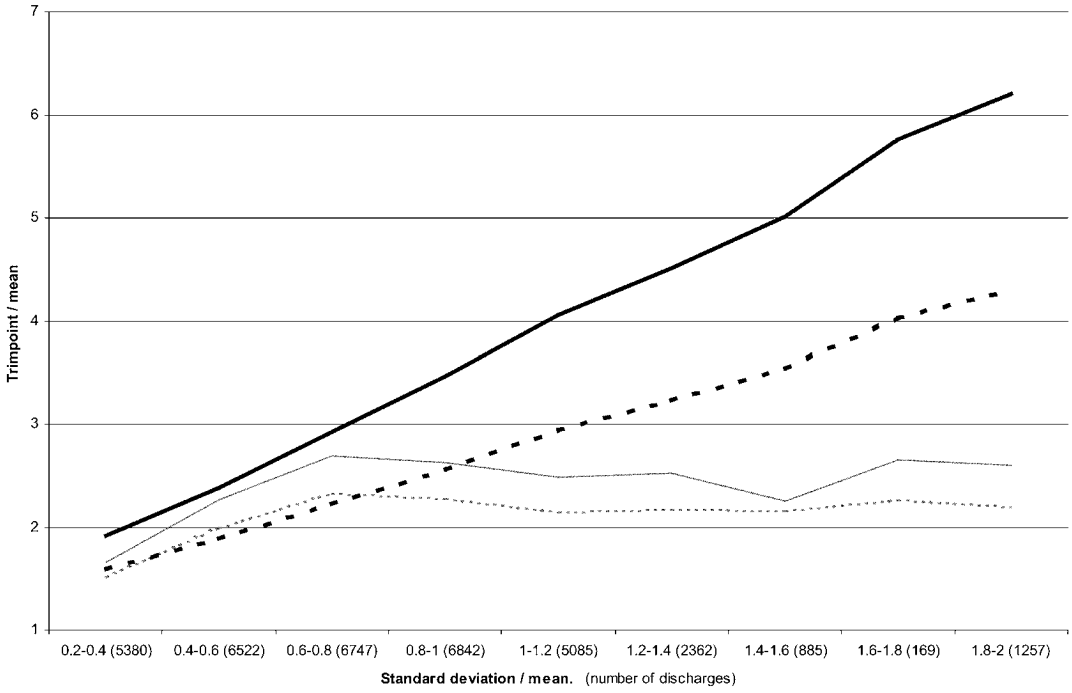

Figure 2. Relationship between trimpoint and intra-DRG variability. (· · ·) Inter-quartile range with parameter = 1.5 (IQ15), (—) inter-quartile range with parameter = 2 (IQ2), (- - -) geometric mean plus two standard deviations (GM2), and (—) geometric mean plus three standard deviations (GM3).

Table 2
Outliers determined by different trimming methods.[a]

	Cases	Costs (MM Pta.)	Cost above trimpoint	Cases	Length of stay	LOS above trimpoint
Total	35,262	12,794		35,262	312,073	
		Percentage of outliers				
		By costs			*By length of stay*	
IQ15	5.9%	20.22%	8.18%	6.75%	16.29%	6.17%
IQ20	4.23%	16.5%	6.68%	4.97%	12.92%	4.77%
GM2	4.76%	17.91%	5.97%	4.98%	15.74%	5.31%
GM3	2.06%	10.56%	3.17%	2.08%	8.53%	2.53%

[a] IQ15: geometric mean plus two standard deviations; and IQ20: geometric mean plus three standard deviations; GM2: inter-quartile range with parameter = 1.5; GM3: inter-quartile range with parameter = 2.

Table 3
Contingency table: inter-quartile range with parameter = 1.5 (IQ15).

Costs	Length of stay		
	Inliers	Outliers	Total
Inliers	32,299	882	*33,181*
Outliers	645	1,436	*2,081*
Total	*32,944*	*2,318*	*35,262*

Table 4
Contingency table: inter-quartile range with parameter = 2 (IQ20).

Costs	Length of stay		
	Inliers	Outliers	Total
Inliers	33,011	758	*33,769*
Outliers	498	995	*1,493*
Total	*33,509*	*1,753*	*35,262*

3. Results

The behaviour of each trimming method analysed in relation to intra-DRG variability and the trimpoint value are shown in figure 2. Non-parametric methods were less sensitive to increased intra-DRG variability than parametric methods.

Table 2 shows the percentage of extreme cost and LOS cases for each of the methods analysed. The percentages of costs and LOS associated with extreme cases are also shown. Between 2 and 6% of cases were cost outliers with an associated cost of between 11 and 20%. The percentage of LOS outliers was between 2 and 7% with associated LOS between 9 and 17% of total LOS. The difference between GM3, used by Medicare, and the other methods was substantial, GM3 being much more conservative and identifying far fewer cases than the other methods. In contrast, IQ15 detected many more cases than the other methods.

Cost and LOS above the trimpoint are also shown in table 2. Between 3 and 8% of total cost was above the cost trimpoint, and between 2.5 and 6% of LOS was above the LOS trimpoint.

The overall distribution of the contingency tables for the four methods is summarised in tables 3–6. The results of applying the "case by case" agreement tests are presented in table 7 and figure 2.

Kappa's coefficient, Youden's index and the ROC curve analyse the diagonal of the contingency tables. When Kappa's coefficient and Youden's index were used, GM2 gave the best valuation and was the second best when the ROC curve was

used. When Landis and Koch's classification of Kappa's coefficient was used, GM2 and IQ15 showed a substantial degree of agreement. The other methods showed a moderate level of agreement.

When the ROC curve was used, GM3 was the method that left the least amount of space below the curve but when other tests were used it gave the greatest values. McNemar's test analyses the symmetry of the contingency table. Only GM3 did not present systematic bias. GM2 presented a value of $\chi^2 = 5.62$, very close to the critical point of this distribution, which was 3.481, with one degree of freedom and 5% significance. In contrast, the non-parametric methods (IQ15, IQ20) gave values that were both very high and far removed from the critical point.

The quality of agreement was evaluated by analysis of extreme cost volume associated with LOS outlier cases. Table 8 shows this analysis for the four methods used. The parametric method with two standard deviations (GM2) revealed the highest percentage (68%) of outlier costs when applied to LOS. The remaining methods revealed a lower percentage of costs, although differences were no higher than 7 points. The value of extreme cost of the false positives revealed by LOS approximation represented 20% of total outlier cost according to GM2. Consequently, when GM2 was used the final overall value detected by the LOS trimpoint was 88%. For the other methods, the final percentage was 80% or less.

The results of the tests used demonstrate that GM2 was the trimming method showing the highest level of agreement between cost and LOS variables, both in terms of number of cases and in terms of the cost revealed.

Unlike the non-parametric methods, the parametric methods did not show systematic bias where non-agreement was found (false positives and false negatives).

Table 5
Contingency table: geometric mean plus two standard deviations (GM2).

Costs	Length of stay		
	Inliers	Outliers	Total
Inliers	32,976	609	*33,585*
Outliers	529	1,148	*1,677*
Total	*33,505*	*1,757*	*35,262*

Table 6
Contingency table: geometric mean plus three standard deviations (GM3).

Costs	Length of stay		
	Inliers	Outliers	Total
Inlies	34,235	301	*34,536*
Outliers	293	433	*726*
Total	*34,528*	*734*	*35,262*

Table 7
Analysis of agreement by trimming methods.[a]

	Kappa	ROC curve	Youden	Sensivity	Specificity	McNemar
GM2	0.64 (0.64–0.67)	0.83 (0.82–0.85)	0.67 (0.64–0.69)	0.69 (0.66–0.71)	0.98 (0.98–0.98)	5.24
GM3	0.58 (0.55–0.61)	0.79 (0.77–0.82)	0.59 (0.55–0.63)	0.60 (0.56–0.63)	0.99 (0.99–0.99)	0.10
IQ15	0.63 (0.61–0.64)	0.83 (0.82–0.84)	0.66 (0.64–0.68)	0.69 (0.67–0.71)	0.97 (0.97–0.98)	36.78
IQ20	0.59 (0.58–0.62)	0.82 (0.81–0.84)	0.64 (0.62–0.67)	0.67 (0.64–0.69)	0.98 (0.98–0.98)	53.82

[a] GM2: geometric mean plus two standard deviations; GM3: geometric mean plus three standard deviations; IQ15: inter-quartile range with parameter = 1.5; and IQ20: inter-quartile range with parameter = 2.
95% confidence interval in parentheses.

Table 8
Extreme costs revealed by application of trimming methods over LOS.

	GM2	GM3	IQ15	IQ20
Application to costs				
Extreme costs (% total cost) (1)	17.9%	10.6%	20.2%	16.5%
Application to LOS				
Associated costs to agreed cases	12.1%	6.2%	13.2%	10.1%
Associated costs to false positives	3.7%	2.3%	3.1%	2.8%
Extreme costs (% total cost) (2)	15.8%	8.5%	16.3%	12.9%
Percentage of revealed cost (2)/(1)	87.9% (84.6–92.3)	80.8% (74.6–84.8)	80.6% (77.0–84.0)	78.3% (74.1–82.3)

[a] GM2: geometric mean plus two standard deviations; GM3: geometric mean plus three standard deviations; IQ15: inter-quartile range with parameter = 1.5; and IQ20: inter-quartile range with parameter = 2.
95% confidence interval in parentheses.

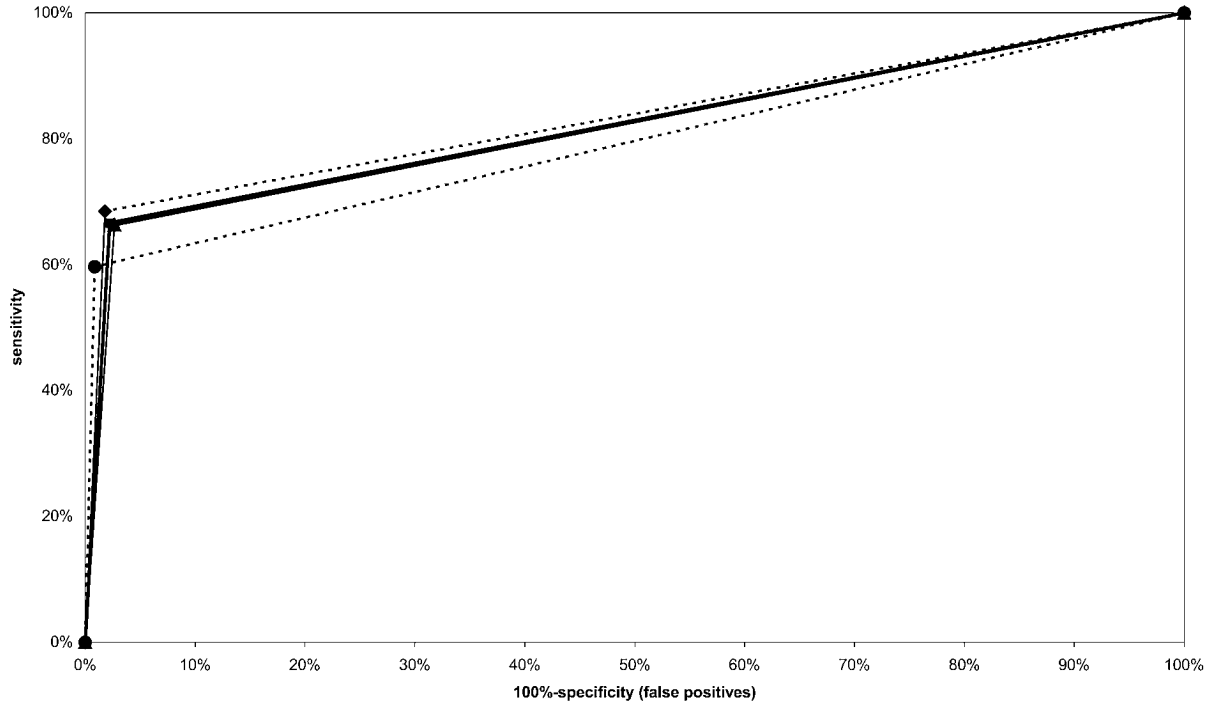

Figure 3. ROC curve for the four trimming methods. (♦) geometric mean plus two standard deviations (GM2), (▲) inter-quartile range with parameter = 1.5 (IQ15), (●) geometric mean plus three standard deviations (GM3) and (■) inter-quartile range with parameter = 2 (IQ2).

4. Discussion

Application of the different trimming methods provides significant information. The correct identification of a small number of cases (between 2 and 5%) would allow the management of between 10 and 20% of resource use. These trimming methods reveal that between 3 and 8% of total costs represents a direct financial risk for the providers of a hospital prospective payment system that does not recognise the existence extreme cases of resource use.

When intra-DRG variability rises the difficulty of determining which trimpoint should be applied also increases; this phenomenon should be incorporated by the trimming method chosen. Non-parametric methods showed a low degree of correlation between trimpoint values and intra-DRG variability, especially when increases in variability were large. In contrast, parametric methods showed a high degree of correlation between trimpoint values and intra-DRG variability. GM2 represents a good compromise because its trimpoint values were similar to those of the non parametric methods when intra-DRG variability was low and showed a high degree of correlation with increases in intra-DRG variability. GM2 was the most satisfactory method used both in the analysis of number of cases and in the analysis of value. When applied to LOS, this method revealed 88% of extreme costs when false positive were included. This is a high figure, allowing effective management of outliers, which grouping of patients into DRGs fails to take into consideration.

GM2 does not reduce the number of extreme cases to the minimum expression and consequently, it may not be the most appropriate method for the complementary funding of these cases. Medicare uses the most conservative trimming method to reduce the percentage of outlier funding to 5% of total expenses for inpatient treatment. Nor is GM2 the method that detects the greatest number of cases and consequently it does not comply with the criteria put forward by Söderlund [33] for choosing IQ15 for the National Health Service in the UK. In this case, the argument for choosing the method that detects the greatest number of cases, costs and LOS is that which makes the resulting mean value for Health Resources Groups more robust.

Therefore, with the hypothesis used in the present study, the most satisfactory method is that which shows the greatest agreement between cost outlier cases and their determination through their application to LOS.

The aim of this type of analysis is to determine cost variability and to identify cases of extreme cost. Thus, it could be argued that:

(1) A small number of cases represent an extremely high percentage of the total cost of a given hospital.

(2) Criteria for choosing the best LOS trimming method can be established according to the method's capacity to predict cost outliers.

(3) In the absence of per patient cost information, the most satisfactory method among those studied was GM2.

(4) Trimming methods and their management are necessary to determine hospital cost variability once this has been adjusted for case mix.

The relevance of the correct determination of outliers to the aspect of health policy that concerns cost variation per patient centres on the need of any analysis attempting to com-

pare inter-hospital costs to correct for outliers, this need being maximal if conclusions about relative efficiency are to be drawn.

Although not the objective of our analysis, if we extend the scope of health policy implications to resource allocation or to hospital services purchasing, our results argue in favour of the measures applied by Medicare's payment system. This organisation allocates part of its budget to fund the extreme cost observed, but constrains the final amount to a marginal level that does not question its DRG-Prospective Payment System. Between 3 and 8% of cost was valued above the trimpoint. This represents the amount of extreme cost that DRGs are unable to value correctly. Application of GM2 revealed that 6% of total inpatient cost would not be correctly financed by any hospital payment system based on case mix techniques unless the system recognised the financial risk generated by outliers [12].

A limitation of our analysis is that it is based on data from two hospitals and therefore, the results cannot be generalised. Even so, application of the GM2 trimming method to the total number of discharges from the Catalan Minimum Data Set (600,000 discharges) between 1996 and 1998 revealed that the percentage of LOS outlier cases and their associated LOS was similar. The cases detected represented 4.49%, a percentage that was very similar to the 4.98% revealed by our analysis. The two hospitals analysed are general hospitals that provide all specialities and that have a case mix index slightly above the average case mix index of the Catalan health system. Therefore, they are highly representative of Catalan case mix; the income per capita in this region is close to the European Union average.

Differences in the percentage of outliers exist between the three levels of case mix complexity in the hospitals of the Catalan public health system. The number of outliers was found to increase with the level of complexity [15]. Thus, any extra payments for outliers should recognise the effect of case mix complexity on the number of outliers; the effect on costs should be analysed in future research projects with a greater number of hospitals for which per-patient cost information is available.

In conclusion, a generalised per-patient cost accounting system in Europe is a utopia that is unlikely to become reality in the short or medium term. Consequently, hospital cost analyses will probably continue to be based on an analysis of LOS consumption adjusted by DRG [2]. Based on the database analysed in the present study, the correlation between costs and LOS is 74%. Outlier cases lead to overvaluation of the estimated mean cost of each DRG due to the lognormal distribution of cost function. For cost analysis, determination of the effect of this overvaluation is indispensable and consequently, the most appropriate trimming method should be used. This is essential because each method yields different results. Thus, criteria for the determination of the most appropriate trimming method have been elaborated and their use is necessary to enable comparison of inter-hospital costs, once these have been adjusted by case mix techniques.

Acknowledgements

The authors wish to thank Mercé Casas for suggesting the subject of the relevance of outlier cases and the quality of trimming methods as a research topic. The authors also wish to thank Montse Rué and Jaume Puig for providing helpful comments on earlier drafts of the paper. This research was supported by the "Fondo de Investigaciones Sanitarias, Instituto Carlos III, Ministerio de Sanidad y Consumo". Project number 1351/1996.

References

[1] P. Armitage and G. Berry, *Statistical Methods in Medical Research* (Blackwell Science Ltd., 1994).
[2] C. Beaver, Y. Zhao, S. McDermid and D. Hindle, Casemix-based funding of northern territory public hospitals: Adjusting for severity and socio-economic variations, Health Economics 7 (1998) 53–61.
[3] A. Briggs and A. Gray, The distribution of health care costs and their statistical analysis for economic evaluation, Journal of Health Services Research and Policy 3 (1998) 233–245.
[4] J.R.G. Butler, *Hospital Cost Analysis* (Kluwer Academic Publishers, The Netherlands, 1995).
[5] J.R.G. Butler, C.M. Furnival and R.F.G. Hart, Estimating treatment cost functions for progressive disease: A multiproduct approach with an application to breast cancer, Journal of Health Economics 14 (1995) 361–385.
[6] K.A. Calore and L. Iezzoni, Disease staging and PMCs. Can they improve DRGs? Medical Care 25 (1987) 724–737.
[7] G.M. Carter, P.D. Jacobson, G.F. Kominski and M.J. Perry, Use of diagnosis-related groups by non-Medicare payers, Health Care Financing Review 16 (1994) 127–158.
[8] G.M. Carter and J.D. Rumpel, *Payment Rates for Unusual Medicare Hospital Cases* (RAND, Santa Mónica, 1992).
[9] M. Casas, Sistemes d'informació hospitalària basats en la casuística: Grups relacionats amb el diagnòstic, Impacte en la gestió hospitalària, Thesis/Dissertation (Universitat Autònoma de Barcelona, Barcelona, 1990).
[10] M. Casas and R. Tomas, Producing DRG statistics at the European level: Lessons from the EURODRG Project, in: *Diagnosis Related Groups in Europe. Uses and Perspectives*, eds. M. Casas and M. Wiley (Springer, Berlin, 1993).
[11] I.R. Chandler, R.B. Fetter and R.C. Newbold, Cost accounting and budgeting, in: *DRGs. Their Design and Development*, eds. R.B. Fetter, D.A. Brand and D. Gamache (Ann Arbor, Michigan, 1991).
[12] F. Cots and X. Castells, Cómo pagamos a nuestros hospitales. La referencia de Cataluña y el contrapunto desde Andalucía, Gaceta Sanitaria 15 (2001) 172–181.
[13] F. Cots, X. Castells, A. Garcia and M. Saez, Relación de los costes directos de hospitalización con la duración de la estancia, Gaceta Sanitaria 11 (1997) 287–295.
[14] F. Cots, D. Elvira, X. Castells and E. Dalmau, Medicare's DRG-Weights in a European environment: The Spanish experience, Health Policy 51 (2000) 31–47.
[15] F. Cots, L. Mercadé and X. Castells, El outlier es un problema estadístico, clínico o asistencial? Red pública de hospitales de Catalunya 1996, Gaceta Sanitaria 14(suplemento 1) (2000) 12.
[16] A. Elixhauser, C. Steiner, R. Harris and R.M. Coffey, Comorbidity measures for use with administrative data, Medical Care 36 (1998) 8–27.
[17] A.R. Feinstein, Clinical epidemiology, *The Architecture of Clinical Research* (Sounders Company, 1985).
[18] R.B. Fetter and J.L. Freeman, Grupos relacionados con el diagnóstico: Gestión por líneas de productos en los hospitales, in: *Los Grupos*

Relacionados por el Diagnóstico: Experiencia y Perspectivas de Utilización, ed. M. Casas (Masson/ S.G. Editores, Barcelona, 1991).

[19] L.I. Iezzoni, A.S. Ash, M. Shwartz, J. Daley, J.S. Hugues and Y.D. Mackiernan, Judging hospitals by severity-adjusted mortality rates: The influence of the severity-adjustment method, American Journal of Public Health 86 (1996) 1379–1387.

[20] L.I. Iezzoni, A.S. Ash, M. Shwartz, B.E. Landon and Y.D. Mackiernan, Predicting in-hospital deaths from coronary artery bypass graft surgery. Do different severity measures give different predictions? Medical Care 36 (1998) 28–39.

[21] E.B. Keeler, G.M. Carter and S. Trude, Insurance aspects of DRG outlier payments, Journal of Health Economics 7 (1988) 193–214.

[22] D.A. Kenny, *Statistics for the Social and Behavioral Sciences* (Little, Brown and Company, Boston, 1987).

[23] H. Lee, Outliers in business surveys, in: *Business Survey Methods*, eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott (Wiley, New York, 1995).

[24] L.K. Lichtig, *Hospital Information Systems for Casemix Management* (Wiley, New York, 1986).

[25] G. López Casasnovas and A. Wagstaff, La financiación hospitalaria basada en la actividad en sistemas sanitarios públicos, regulación de tarifas y eficiencia: El caso de la concentración hospitalaria en Cataluña, in: *La Regulación de los Servicios Sanitarios en España*, eds. G. López Casasnovas and D. Rodríguez (Editorial Civitas, Madrid, 1997).

[26] J.P. Newhouse, Reimbursing health plans and health providers: Efficiency in production versus selection, Journal of Economic Literature 34 (1996) 1236–1263.

[27] S. Peiró, Limitaciones en la medición de los resultados de la atención hospitalaria: Implicaciones para la gestión, in: *Instrumentos Para la Gestión en Sanidad*, ed. AES (SG-Editores S.A., Barcelona, 1995).

[28] G. Rhodes, M. Wiley, R. Tomas, M. Casas and R. Leidl, Comparing EU hospital efficiency using diagnosis-related groups, European Journal of Public Health 7 (1997) 42–50.

[29] D.L. Sackett, R.B. Haynes, G.H. Guyatt and P. Tugwell, Epidemiología clínica, in: *Ciencia Básica para la Medicina Clínica* (Panamericana, Buenos Aires, 1994).

[30] H.F. Sanderson, P. Anthony and L.M. Mountney, Healthcare resource groups. Version 2, Journal of Public Health Medicine 17 (1995) 349–354.

[31] T. Smith, D. Archer and F. Butler, Healthcare resources groups and general practitioner purchasing, in: *Casemix for All*, eds. H. Sanderson, P. Anthony and L. Mountney (Radcliffe Medical Press, Oxon, 1998).

[32] N. Söderlund, Product definition for healthcare contracting: An overview of approaches to measuring hospital output with reference to the UK internal market, Journal of Epidemiology Community Health 48 (1994) 224–231.

[33] N. Söderlund, A. Gray, R. Milne and J. Raftery, The construction of resource-weights for healthcare resource groups: A comparison of alternative data sources and methodological approaches (National Casemix Office, London, 1996).

[34] N. Söderlund, A. Gray, R. Milne and J. Raftery, Case mix measurement in english hospitals: An evaluation of five methods for predicting resource use, Journal of Health Services Research and Policy 1 (1996) 10–19.

[35] J.L. Temes, J.L. Díaz and B. Parra, *El Coste por Proceso Hospitalario* (Editorial Interamericana McGraw-Hill, Madrid, 1994).

[36] J.W. Thomas and M.L.F. Ashcraft, Measuring severity of illness: Six severity systems and their ability to explain cost variations, Inquiry 28 (1991) 39–55.

[37] S. Udpa, Activity-based costing for hospitals, Health Care Management Review 21 (1996) 83–96.

[38] D.W. Young and L.K. Pearlman, Managing the stages of hospital cost accounting, Healthcare Financial Management 47 (1993) 58–80.