ELSEVIER

European Journal of Cardio-thoracic Surgery 29 (2006) 431—433

EUROPEAN JOURNAL OF CARDIO-THORACIC SURGERY

www.elsevier.com/locate/ejcts

MADCAP: a graphical method for assessing risk scoring systems

Steve Gallivan [a,*], Martin Utley [a], Domenico Pagano [b], Tom Treasure [c]

[a] Clinical Operational Research Unit, University College London, UK
[b] University Hospital Birmingham, UK
[c] Guy's and St. Thomas's Hospitals Medical School, London, UK

Received 7 October 2005; received in revised form 19 December 2005; accepted 21 December 2005; Available online 17 February 2006

Abstract

Objective: We set out to develop a method for assessing the performance of clinical risk models over the spectrum of risks and to assess the performance of the EuroSCORE risk model used in cardiac surgery. **Methods:** We developed a graphical method for assessing the performance of clinical risk models over the spectrum of risks. To illustrate the technique, we analysed retrospective data concerning 9268 patients that underwent cardiac surgery and for whom both the additive EuroSCORE prediction of risk of morality and vital status at 30 days were available. **Results:** The graphical tool developed, called MADCAP (Mean Adjusted Deaths Compared Against Predictions), can be used to highlight systematic features of the performance of a clinical risk model. Its use in the current study indicates that the additive version of the EuroSCORE model seems to underestimate risk amongst low-risk cases (0% and 1%). Otherwise the score systematically favours risk avoiding behaviour as the risk model underestimates mortality for 2—6% prediction but not at 7% and above. **Conclusion:** The robustness of case-mix adjusted audit is dependent on the performance of the risk scoring system over the entire spectrum of risk. If we are to use risk adjustment of mortality rates when comparing outcomes obtained by different units or individual surgeons, it is essential that we continually review the performance of the risk adjustment method. The MADCAP method presented here provides a useful tool to this end.

Keywords: Risk models; Audit; Cardiac surgery; Mortality

1. Introduction

When auditing perioperative deaths, it is necessary to take case mix into account, otherwise surgeons who take on the more demanding and complex cases might be unfairly assessed if their mortality rate seems atypically high. In view of this, there has been a growing interest in the topic of risk scoring whereby preoperative factors such as age and clinical status are used to forecast the intrinsic probability of a perioperative death. In cardiac surgery, two scoring systems are commonly used, one due to Parsonnet et al. [1] and a more recent development, the EuroSCORE [2]. Such scoring systems are also used in other clinical contexts [3,4].

In assessing the merits of a scoring system, it is important that it should be seen to give reasonably good predictions across the whole spectrum of cases that are typically encountered. If there is systematic under- or overestimation of risk for any part of the risk spectrum, this potentially degrades the audit process and indeed may indirectly promote poor practice whereby cases are avoided if the notional risk score falls short of the truth.

Here we describe a simple graphical method that can be used in the assessment of risk scoring systems. We apply it in the case of the additive version of the EuroSCORE highlighting systematic biases.

2. Methods

We first rank cases in order of risk according to the forecasts of the risk model being scrutinised. Using this ordering, two cumulative graphs are charted. One accumulates the forecasted probabilities of death and thus gives a running tally of the number of deaths that would be expected under the assumptions of the risk model. In contrast to this, we plot a cumulative sum of the deaths that actually occurred. In the simplest instance, where each case had a unique risk score and there are no duplicates, this would simply be a graph that steps one unit across for each case and up for each death. If the risk scoring method is flawed, systematic divergence of the two graphs highlights discrepancies. Systematic divergence is further illuminated by examining a graph of the arithmetic difference between the two plots, a technique similar that used in a VLAD plot [5].

There is an irksome complication. In additive EuroSCORE, more than 95% of cases in a typical case mix are scored in

* Corresponding author. Fax: +44 207 813 2814.
E-mail address: s.gallivan@ucl.ac.uk (S. Gallivan).

1010-7940/$ — see front matter
doi:10.1016/j.ejcts.2005.12.057

integers from 0 to 10, so tied scores are frequent. This poses the problem of how to represent the cumulative plot of actual deaths. The ordering chosen for large clusters of cases of equal risk could give a chart with upward steps in different places while the visual impression of 'goodness of fit' is crucially dependent on the choice of where these steps occur. This raises the possibility of accidental (or deliberate) bias whereby cases might be ordered in such a way so as to give a misleading impression of good or bad fit.

We resolve this problem by assuming that all possible orderings of cases with equal predicted risk are equally valid. Under this assumption, an unbiased representation of the outcome data is given by simply taking the average of all possible orderings. To describe this process, we have coined the acronym MADCAP (Mean Adjusted Deaths Compared Against Predictions). Even with moderately sized data sets, there may be numerous patients with tied risk scores and thus an alarmingly high number of different ways of ordering the cases. Fortunately, a simple method is sufficient to resolve this difficulty. Using straight lines in the cumulative graph of actual mortality gives an unbiased representation of deaths within groups of patients with tied risk scores (see Appendix A).

To illustrate the use of the MADCAP method, we use data prospectively collected in two large units in England comprising a case by case record of EuroSCOREs and outcome (in-hospital mortality). The data sets were anonymised for both patients and surgeons and were merged for analysis.

3. Results

The MADCAP chart for the 9268 cases is shown in Fig. 1 and the arithmetic difference between the two traces is shown in Fig. 2, which highlights any systematic bias. This figure also illustrates the risk score bands into which the data fell.

4. Discussion

When interpreting the charts produced by the MADCAP technique introduced in this paper, it is important to realise

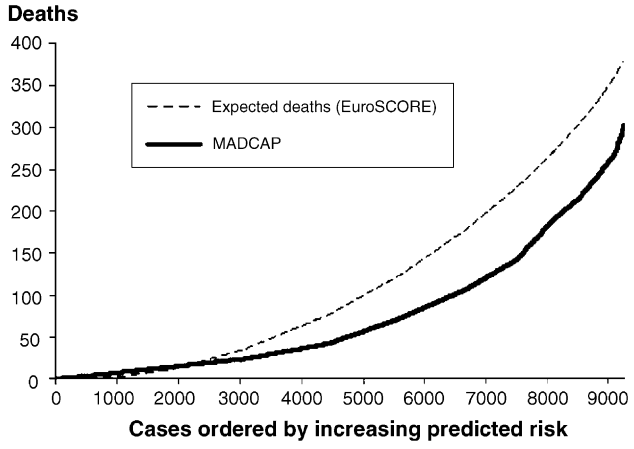

Fig. 1. The MADCAP (Mean Adjusted Deaths Compared Against Predictions) chart for 9268 cardiac surgery cases comparing expected and actual cumulative mortality.

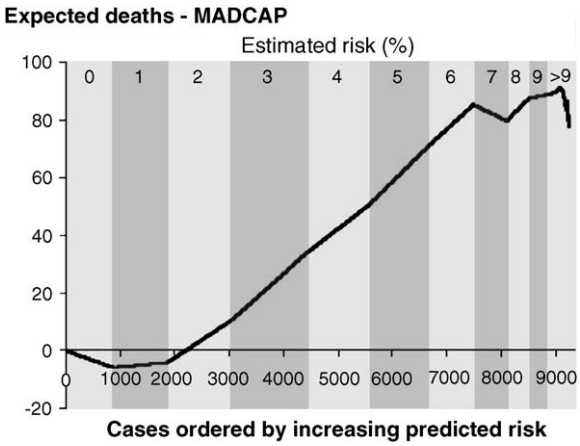

Fig. 2. The difference between expected and actual cumulative mortality for cases ordered by increasing risk using the MADCAP (Mean Adjusted Deaths Compared Against Predictions) method.

that some chance discrepancy between actual and expected mortality would be likely even if the risk that each patient faced were exactly as forecast by the risk model. The MADCAP is a visual aid for distinguishing between such chance discrepancy and systematic failings in the performance of a risk model. Although by no means a statistical measure of 'goodness of fit', it is suitable for use in the development of risk models [6] as well as for assessing the performance of an established risk model. The MADCAP technique is applicable whether the risk model concerned is derived from multiple regression analysis or any other method and we recommend the routine use of this check for systematic bias in the process of developing a risk model.

The technique is intended to be part of the research methodology of those engaged in the development or evaluation of risk models rather than part of day to day clinical practice. However, as such risk models are becoming an increasingly common part of audit and communication with patients, it is important that all clinicians have an appreciation of how accurate such risk models are and of their strengths and weaknesses.

With regards to the data presented in Figs. 1 and 2, it would seem that there are systematic discrepancies between the actual mortality amongst the cases studied and that expected according to the additive EuroSCORE risk model. Such discrepancies are not exposed by use of the ROC curve (see for example [7]).

Fig. 1 shows that mortality was lower than predicted overall. Fig. 2 shows that mortality was greater than predicted amongst low-risk cases (0% and 1%) and perhaps also amongst high-risk patients ($\geq$7%).

It seems self-evident that prediction of 0% mortality provides an unrealistic quality target. In this range, average mortality is always likely to be higher than the prediction since it cannot be below it. The same argument could be made for all predictions $\leq$1% and yet 20% of the cases are scored at 0% or 1% risk. Loading the practice of 'beginners', for example with these cases and then comparing the risk adjusted results with surgeons with a wider case mix could suggest apparently less good results.

Apart from the very low-risk cases, the score systematically favours risk avoiding behaviour as the risk model underestimates mortality for 2—6% prediction but not at 7% and above (Fig. 2). This has been noted previously [8] but in an analysis depending on already grouped data sets. This failure to reflect increasing risk accurately is unfair on surgeons taking on the very patients whose lives are most at risk from the natural history of the disease since most elements in the perioperative risk score are also markers for risk of death without surgery. These are the patients who have the most to gain from heart surgery because it is where surgery makes the biggest difference, and a scoring system, which rewards risk averse case selection is not in the patients' interests. If we are to use risk adjusted death rates as a comparative index of performance, it is essential that we continually review the performance of the risk adjustment method. The MADCAP charts presented in this paper provide a useful tool to this end.

The technique described in this paper is designed purely to detect systematic flaws in a risk model. For instance, had the method we suggest been used when the additive EuroSCORE was developed, the systematic bias that is a feature of this scoring system would have been apparent. To make recommendations regarding how to improve a risk model, if indeed it is judged that improvements are required, is beyond the scope of our current work. That said, one option would be to use the discrepancies highlighted by the use of MADCAP to adjust the risk score. This would have the advantage of making the model better, according to the criteria of the visual appearance of the MADCAP chart, but may well have disadvantages in terms of other criteria. A MADCAP chart on its own is not sufficient to judge whether a risk model is good or bad. The acid test is to consider what uses the risk model will have and whether it is fit for these purposes.

Acknowledgements

The authors would like to thank Ben Bridgewater and colleagues at Wythenshawe Hospital Manchester and D.P.'s colleagues at University Hospital Birmingham.

References

[1] Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation 1989;79(Suppl. I):I3—12.
[2] Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (*Euro*SCORE). Eur J Cardiothorac Surg 1999;16:9—13.
[3] Lawrance RA, Dorsch MF, Sapsford RJ, Mackintosh AF, Greenwood DC, Jackson B, Morrell C, Robinson MB, Hall AS. Use of cumulative mortality data in patients with acute myocardial infarction for early detection of variation in clinical practice: observational study. Br Med J 2001;323:324—7.
[4] Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. Br J Surg 1991;78(3):355—60.
[5] Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the result of cardiac surgery by variable life adjusted display (VLAD). Lancet 1997;350:1128—30.
[6] Berrisford R, Brunelli A, Rocco G, Treasure T, Utley M. Audit and guidelines committee of the European Society of Thoracic Surgeons; European Association of Cardiothoracic Surgeons. The European Thoracic Surgery Database project: modelling the risk of in-hospital death following lung resection. Eur J Cardiothorac Surg 2005;28:306—11.
[7] Michel P, Roques F, Nashef SAM. Logistic or additive EuroSCORE for high-risk patients? Eur J Cardiothorac Surg 2003;23:684—7.
[8] Gogbashian A, Sedrakyan A, Treasure T. EuroSCORE: a systematic review of international performance. Eur J Cardiothorac Surg 2004;25:695—700.

Appendix A. The equivalence of the MADCAP calculations to linear interpolation

Suppose one is concerned with a data set with clusters of cases, operations within each cluster having a tied risk score. For different permutations of the order of cases, a different cumulative mortality graph might result. It will be shown that by taking the mean of all such cumulative mortalities, over all possible permutations of cases in the cluster, gives a linear increasing function. In view of this, amalgamating such cumulative plots over all the clusters gives a piecewise linear graph expressing the mean cumulative mortality for the whole dataset. This is the basis of the MADCAP charting method.

The proof of this is an example where the intuitively obvious (linear interpolation) is troublesome to justify mathematically.

Consider a single cluster of N cases with tied risk score for which there have been D deaths.

There are $N!$ possible permutations of the cases, each of which could be used to construct a cumulative chart of mortality. Although requiring some thought, it can be seen that for $1 \leq i \leq N$, the proportion of permutations within the cluster that have a death assigned to the ith case is D/N and for these, the cumulative mortality increases by 1, while there is no increase for other permutations. Thus at the ith step, the mean increase in cumulative mortality is D/N, $1 \leq i \leq N$.