STATISTICALLY SPEAKING

Correspondence to:
Professor Jaideep J Pandit
St John's College
Oxford OX1 3JP
Tel: 01865 221590
Fax: 01865 220027
email: jaideep.pandit@dpag.ox.ac.uk

Dispassionate indicator or evil curse: are some scoring systems for predicting postoperative mortality lethal?

J.J. Pandit

Consultant Anaesthetist, Nuffield Department of Anaesthetics, Oxford University Hospitals NHS Foundation Trust, Oxford OX3 9DU

Several scoring systems have been used to predict mortality from major surgery. By predicting risk in this way, the score can help select patients for surgery, or identify those who need specific interventions to manage their risk [1,2]. Hitherto, it has not been suggested that the very act of using a score either actually harms patients (akin to a curse) or protects them (like a holy incantation). However, some recent data suggests that the former may be the case: the act of performing a predictive score can be lethal, *even if this score does not influence subsequent management*. How can this be, and what is the evidence?

Carlisle recently commented in an editorial [3] on data from de Buck van Overstraeten et al., assessing the value of using the CR-POSSUM score in a Belgian cohort [4]. The original study was well conducted with interesting conclusions, but those aspects are not our concern here. In his editorial, Carlisle statistically analysed the results of applying three other common scores (SORT [5], the UK audit model [6] and the Torbay model [7]) to the Belgian data, in addition to the CR-POSSUM score [8] used by de Buck van Overstraeten et al. Note that in this theoretical analysis, the management of patients was not dictated by the score outcome, which was something only retrospectively applied at a distance. Carlisle's aim was to show that some predictive models (he concluded the UK audit tool and Torbay

model) were more consistent with the actual mortality rates published in the Belgian study. However, the same data can be presented in different ways to address distinct issues. I was interested in a completely separate question: what can be learned from plotting the mortalities predicted by the various scoring systems?

So, I have plotted the same data used by Carlisle in a contrasting way (Figure 1). The x-axis of Figure 1 (non-linear scale) represents the one month mortality predicted by each of the four scoring systems for the 11 patients in the Belgian cohort who died. The y-axis is the cumulative mortality of this group. Thus for each scoring system the lowest-scoring patient is plotted first (lowest) on the y-axis, and each death represents ~9.1% of the total deaths. I have then fitted a simple sigmoidal best fit relationship to each (least squares non-linear regression, Sigmaplot for Windows Version 11.0, Systat Software Inc, CA, USA).

The result is something of a shock. Readers might immediately recognise that the graph is rather like a set of drug dose-response relationships, where each scoring system can be regarded conceptually as a drug that is 'administered', and the resulting 'drug effect' on mortality presented on the y-axis. The position of the 'drug' relationship on the x-axis represents its potency (toxicity). Clearly the most potent (toxic or harmful) 'drug' is that which lies to the left of the graph, and the weakest (safest) is that which lies to the right. The shock is that the very act of applying some scoring systems (here, the SORT) is lethal - *even when this score is itself not influencing subsequent management* - whereas the act of applying others (here, the CR-POSSUM) is potentially protective

To explain the graph a little more, predicting a mortality of 5% (0.05 on x-axis) through verbally administering SORT at the bedside would result in a cohort mortality of 100%. At the other extreme, predicting exactly the same mortality of 5% through incanting CR-POSSUM instead over the patient's bed would result in nobody dying. In between these, mortality predictions of 5% result in actual cohort death rates of ~80% for the UK audit model and just ~20% for the Torbay model (Figure 1).

The frightening notion that actions unrelated to management or therapy may be highly effective (or harmful) is not new. Counsel et al. reported on the miraculous effect of simply rolling a dice on stroke outcomes [9]. Prayer has been reported as highly effective [10] - and also harmful [11], where the recitation is possibly a curse. Anaesthetists may have hitherto regarded the scoring systems available to us as highly sophisticated, quantitative and

objective tools. Yet Figure 1 demonstrates that there may be a direct effect of intoning the scoring system itself. Clearly, we should be much more apprehensive about which scoring system we employ, and what we predict remotely or at the bedside. It might be lethal.

Or: is the result in Figure 1 simply an illusion that teaches us to be more careful about how we plot our data when analysing the utility of predictive scoring systems? Especially concerned readers should quickly move on to the discussion presented as in the Appendix.

Acknowledgement and Competing interests

JJP is Editor of Anaesthesia, and this article has undergone additional external review. I thank Dr John Carlisle for sharing his analysis and for Dr de Buck van Overstraeten et al. for sharing their data to enable Figure 1 to be re-plotted. Supplementary material is provided to reassure readers who may be distressed at the contents of this article.

References

1. Carlisle JB. Simulations of the effects of scheduled abdominal aortic aneurysm repair on survival. *Anaesthesia* 2015; **70**: 666–78.

2. Howell SJ. Predicting survival after surgery: a matter of life and death. *Anaesthesia* 2015; **70**: 637–40.

3. Carlisle J. Commentary: ‘Life is a line, not a dot’. *Colorectal Disease* 2017; **19**: 64-66.

4. de Buck van Overstraeten A, Stijns J, Laenen A, Fieuws S, Wolthuis AM, D'Hoore A. Is colorectal surgery beyond the age of 80 still feasible with acceptable mortality? An analysis of the predictive value of CR-POSSUM and life expectancy after hospital discharge. *Colorectal Di*sease 2017; **19**: 58-64

5. Protopapa K, Simpson J, Smith N, Moonesinghe S. Development and validation of the Surgical Outcome Risk Tool (SORT). *British Journal of Surgery* 2014; **101**: 1774–83.

6. Walker K, Finan P, van der Meulen J. Model for risk adjustment of postoperative mortality in patients with colorectal cancer. *British Journal of Surgery* 2015; **102**: 269–80.

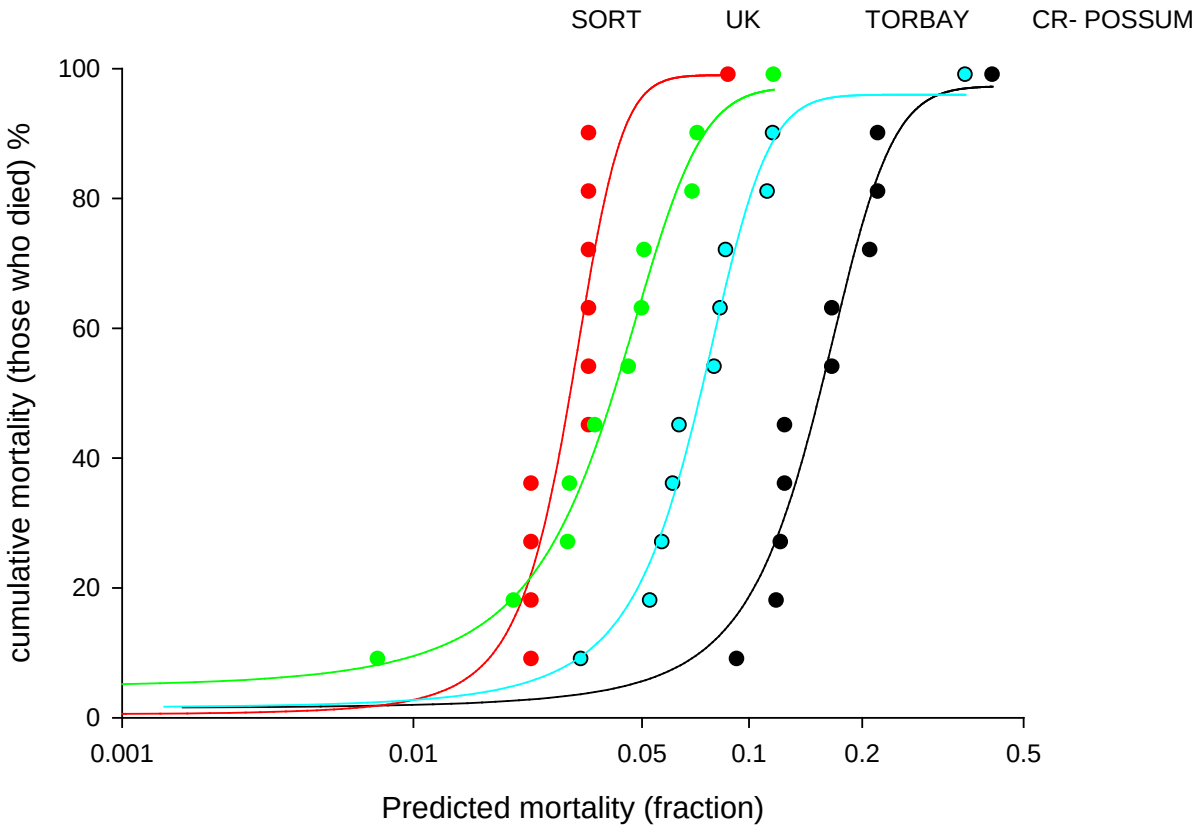
7. Carlisle J, Danjoux G, Kerr K, Snowden C, Swart M. Validation of long-term survival prediction for scheduled abdominal aortic aneurysm repair with an independent calculator using only pre-operative variables. *Anaesthesia* 2015; **70**: 654–65.

8. Tekkis P, Prytherch D, Kocher H et al. Development of a dedicated risk-adjustment scoring system for colorectal surgery (colorectal POSSUM). *British Journal of Surgery* 2004; **91**: 1174–82.
9. Counsell CE, Clarke MJ, Slattery J, Sandercock PA. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *British Medical Journal*1994; **309**: 1677-81.

10. Lesniak KT. The effect of intercessory prayer on wound healing in nonhuman primates. *Alternative Therapies in Health and Medicine* 2006; **12**: 42–8.

11. Benson H, Dusek JA, Sherwood JB, Lam P, Bethea CF, Carpenter W, et al. Study of the Therapeutic Effects of Intercessory Prayer (STEP) in cardiac bypass patients: A multicenter randomized trial of uncertainty and certainty of receiving intercessory prayer. *American Heart Journal* 2006; **151**: 934–42.

Figure 1. Explanation in text. Cumulative mortality (y-axis) for 11 patients in the data cohort analysed by Carlisle [1] based on a Belgian study [2]; plotted against (x-axis) the predicted mortality (decimal fraction) of four scoring systems. The curves are best fit sigmoidal fitted using non-linear least squares regression.



Appendix: Resolving the graphical paradox

If some readers are disturbed by the content of the main accompanying article, it may be important to offer some reassurance. Although the plot presented in Figure 1 is entirely valid, some caution needs to be exercised in its interpretation. The last sentences of the main article hint at the solution to the frightening paradox. The plot presented is rather like a mathematical/statistical optical illusion. The graph is actually one way of plotting the distribution of predicted mortality by different scoring systems, only for patients who die. The cumulative nature of the y-axis allows us to define a median (amongst dead patients only). The separation of the curves suggests a difference in calibration between the predictive scores.

First, Figure 1 clearly shows the limits of quantitative mortality prediction of some scoring systems as applied to this cohort of patients (ie, their calibration). In other words, SORT never predicts a mortality >5%, while for example, CR-POSSUM does not predict any mortalities <10%. The individual plots are therefore artificially constrained, and this creates an inherent bias or skew that helps yield the dramatic message.

Second, the y-axis must be viewed as artificial. It only relates to the patients who died in the cohort and does not include the scores for patients who did not die. Overall mortality was 11/207 (~5.3%) in the original dataset, and not 100% as seemingly indicated on the y-axis. The cumulative nature of the plot creates the effect; the patient with the lowest score for each scoring system is arbitrarily plotted first (lowest) on the x-axis. There is no a priori reason to do this (eg, a different plot might instead have plotted the patient who died first, lowest on the axis). In other words, all the patients presented in the plot unfortunately died. So it makes no sense to say that, for example, that when the Torbay method predicts a mortality of 5% about half the patients die – because in fact they all died.

Third, for any given patient each scoring system predicts a different mortality (Figure 2). Thus, the highest scoring patient by CR-POSSUM (who is then plotted at the very top of the y-axis in Figure 1 on the black line and symbols) is in fact one of the lower-scoring patients by SORT (and thus plotted low on the red line and symbols in Figure 1). This adds to the artifice of the shocking result of Figure 1 (ie, when we look across from any point on the y-axis, we are not looking at the same patient). Moreover as Carlisle concludes in his editorial [1], the middle two methods (UK audit tool and Torbay model) yield actual mortality data that better fit the data of de Buck van Overstraeten et al. [2]. The shock effect is created by the outlier scoring systems.

Finally, as Carlisle stresses in his editorial, all data are based on 30-day mortality, which itself is an artifice.

The above counter-points should all reassure that the plot of Figure 1 is nuanced. It cannot be the case that a remote incantation influences mortality. However, the case illustrates the care needed when using or interpreting results of scoring systems. Different methods yield quantitatively different 30-day mortality estimates (itself a cut-off arguably irrelevant). Some methods probably work better for some patient populations than others. There are problems with scoring systems, and serial calibration plots over time would better demonstrate that some scores fit some populations better than others. Scoring systems should probably be regularly updated and calibrated to the population of interest in order to predict outcomes more accurately, and allow for appropriate risk-adjustment. Figure 1 underlines the caution needed if planning service delivery based on scoring systems for large patient populations.

Figure 2. The mortality predicted by SORT vs that predicted by CR-POSSUM for the cohort of 11 patients analysed by Carlisle [1]. Note the same patient is predicted different mortalities, and this also applies if the other systems are examined.