Intensive Care Med (2002) 28:105–107
DOI 10.1007/s00134-001-1186-1

EDITORIAL

F. Shann

Are we doing a good job: PRISM, PIM and all that

Accepted: 15 November 2001
Published online: 12 January 2002

F. Shann (✉)
Royal Children's Hospital, Flemington Road, Parkville,
Victoria 3052, Australia
e-mail: shann@cryptic.rch.unimelb.edu.au

It is very important that we monitor how well we are looking after patients admitted to our intensive care units. Indeed, many governments now demand that we do this. It is important that intensivists understand the basic principles of the statistics involved, and the limitations of the techniques [1].

An important way to tell if we are doing a good job in paediatric intensive care is to use a risk-of-mortality model such as PRISM (Paediatric Risk of Mortality) or PIM (Paediatric Index of Mortality) [2, 3]. We count the number of patients who actually died in our unit (observed deaths), and calculate the number of deaths predicted by a model such as PRISM or PIM (expected deaths). We then calculate our standardised mortality ratio (SMR) by dividing the number of observed deaths by the number of expected deaths: an SMR less than 1.00 suggests we are doing better than expected and a number greater than 1.00 suggests we are doing worse than expected. If the 95% confidence intervals of the SMR include 1.00, any variation from 1.00 may well be due to chance. In effect, the SMR compares the number of deaths in the sample (for example, in your unit last year) with an estimate of the number of deaths that would have occurred if the same patients had been looked after in the units that derived the score (at the time the score was derived).

In this issue, Gemke and van Vught have studied the performance of PRISM and PIM in a small sample of children in intensive care in Utrecht [4]. Risk of mortality models such as PRISM and PIM are developed by collecting information about a large number of patients from representative intensive care units, and observing which children die. The information is examined to see which variables predict whether children die or survive. Because the variable we are trying to predict has just two values (death or survival), we use logistic regression analysis to derive an equation that describes the relationship between the predictor variables (like blood pressure and pH) and mortality [5]. (If we were trying to predict a continuous variable, such as the weight of a patient from her age, we would use ordinary least-squares regression, rather than logistic regression.)

One problem with logistic regression is that it is quite hard to tell whether the model really is a good description of the data – whether it is a good way to predict how many children are going to die. There are two tests that look at different aspects of the performance of a logistic regression model. The first calculates the area under the receiver operating characteristic (ROC) plot, which is a graph of the sensitivity versus the specificity for every value of the score; an area of 1.00 suggests a perfect model and 0.50 would be expected by chance. An area of 0.70–0.79 is acceptable, 0.80–0.89 is good, and 0.90 or more is excellent. This test is equivalent to ranking all the scores from worst to best, then comparing the average rank for non-survivors with the average rank for survivors; with a perfect score, all the non-survivors would have a higher rank than all the survivors.

The trouble with the area under the ROC plot is that it does not tell us whether the model predicts mortality well for both ill and not-so-ill children. The Hosmer-Lemeshow test was developed to deal with this problem [5]. Unfortunately, the results of the test are often misinterpreted. All the scores are ranked from best to worst and then divided into ten groups (known as deciles of risk). Within each of the ten groups, the observed number of deaths is compared to the number predicted by the model, and the observed number of survivors is com-

Table 1 Hosmer-Lemeshow table for children in Melbourne in 2001

Group	Predicted probability of death	Died		Survived		Total
		Observed	Expected	Observed	Expected	
1	0.0076	0	0.9	150	149.1	150
2	0.0093	0	0.6	72	71.4	72
3	0.0120	0	0.9	85	84.1	85
4	0.0152	0	1.4	103	101.6	103
5	0.0206	0	1.8	102	100.2	102
6	0.0296	3	2.6	99	99.4	102
7	0.0424	8	3.7	95	99.3	103
8	0.0643	0	5.3	102	96.7	102
9	0.1487	9	9.5	93	92.5	102
10	0.9956	21	32.7	81	69.3	102
Total		41	59.4	982	963.5	1023

pared to the number predicted. Table1 shows the Hosmer-Lemeshow tabulation for children admitted to intensive care in Melbourne in the first 9 months of 2001. Careful analysis of the information in the table is much more valuable than just looking at the *p* value. The rows at the top of the table have been classified as low risk by PIM: there are no deaths, and the numbers of actual and predicted survivors match well on each row (so PIM is predicting well in these low-risk patients). The bottom rows represent high-risk patients: all 41 deaths occur here, and there is a suggestion that, in high-risk patients, our standard of care in 2001 is better than it was when PIM was derived. Overall, PIM predicts too many deaths in Melbourne (41 children died, 59.4 predicted) – mortality prediction models have to be updated regularly as standards of care improve. A new edition of PIM will be available soon.

Too much weight is often given to the Hosmer-Lemeshow *p* value. First, the *p* value is particularly unreliable with sample sizes less than 400, or when there are few deaths, or when more than four of the 20 values in the expected columns in the table are less than 5.0. This is the case with the data from Utrecht, and in these circumstances, the *p* value often changes dramatically with small changes in the data. Secondly, the Hosmer-Lemeshow *p* value is highly unstable when the number of covariate patterns is lower than the number of subjects, as is usually the case with PIM and PRISM data. Bertolini et al. performed the Hosmer-Lemeshow test using all possible subject dispositions on data from 1393 ICU patients – they obtained about one million different *p* values, ranging from 0.01 to 0.95 [6]. Thirdly, the Hosmer-Lemeshow *p* value does not tell us about the clinical importance of a difference between actual and predicted survivors and non-survivors. A small (clinically unimportant) difference in a large sample, and a large (clinically important) difference in a small sample will both give the same *p* value. Just as we should put more emphasis on confidence intervals than on *p* values, we should put more weight on inspection of the Hosmer-Lemeshow table than on whether the *p* value is more or less than 0.05.

When you analyse your data, the Hosmer-Lemeshow test will almost always show a poor fit of the model – unless the numbers of observed and predicted survivors and non-survivors are similar at all deciles of risk (that is, on all ten lines of the table). A significant Hosmer-Lemeshow test does *not* demonstrate that PRISM or PIM is inappropriate for your unit – it is far more likely to occur because the standard of care in your unit is better or worse than in the units that derived the score (at the time it was derived). If your SMR is significantly different from 1.00 (the 95% confidence limits do not include 1.00), you should expect the Hosmer-Lemeshow *p* value to be less than 0.05. Far too often, when the Hosmer-Lemeshow *p* value is less than 0.05, investigators conclude (incorrectly) that this means that the mortality prediction model that they have used is not valid in their intensive care unit. They then derive a new logistic regression model, which will apply only to their unit. This defeats the main purpose of these models – which is to allow you to compare the standard of care in your unit to the standard of care in the derivation units (at the time the model was developed).

It is very important to remember that intensive care mortality models should only be used in groups of patients; they should never be used to guide the management of an individual patient. It is sensible to use PRISM or PIM to compare the actual number of deaths in your unit with the number of deaths predicted by one of the models, but not to decide that an individual patient is too sick to be worth treating. None of the models is anywhere near accurate enough to be used for individual patients. For example, even in the very high risk row at the bottom of the table, 81 of the 102 children survived.

PRISM and PIM are risk-of-mortality models, not severity-of-illness models. It is true that, on average, patients with a low risk of dying are not as sick as those with a high risk. However, this is not true for all types of patient. For example, children with severe croup are very likely to die without intensive care, but they have a very low mortality if they are properly managed in intensive care – so PRISM and PIM give these children a low score despite the fact that they are very ill.

So how do we go about using PRISM or PIM to assess the quality of care in our intensive care unit? First we have to collect the data, and accurate collection of data is difficult and time-consuming, but crucially important. It is important to read the instructions very carefully, and to read them again every few months. The data must be collected by a small number of properly trained people who are interested in the job and who understand the importance of accuracy. Both PRISM and PIM assume that most patients are not very ill (the constant in the equations is negative) and most of the information collected increases the estimated risk of mortality – so you will predict fewer deaths if you fail to record relevant information. Every year or so, a random sample of data from 20–50 of your patients should be collected in duplicate by another person to check the accuracy of data collection.

An adequate number of patients must be studied before your SMR will mean very much – if you have less than about 50 deaths in your sample, the confidence intervals will be so wide that you may not detect important differences in mortality. In addition, if you use an old model, you are comparing your performance to an old standard – for example, the data for PRISM II were collected in the United States between 1980 and 1985, and you should really use a more recent model such as PRISM III or PIM.

Mortality prediction models such as PRISM or PIM are very useful tools if they are used properly. The model is probably appropriate for your unit if the area under the ROC plot is greater than 0.7, and the number of observed and expected survivors (and non-survivors) is similar across all ten rows of the table, or if there is a fairly consistent ratio of observed to expected. On the other hand, the model may not be appropriate if the area under the ROC plot is less than 0.7 or if, for example, your unit has many more deaths (and fewer survivors) than predicted in the low risk groups in the table, but many fewer deaths (and more survivors) in the high risk groups. If the model is appropriate for your unit, the SMR will give you useful information. For example, an SMR of 0.88 (95% CI 0.55–1.20) suggests that the standard of care in Utrecht in 1997 was comparable to the standard of care in Australia and Birmingham between 1994 and 1996 (although the Utrecht study was small and the confidence intervals are therefore rather wide) [4]. It is encouraging to know when our work is up to standard, and it can be very helpful to know that some units or regions are below standard so that steps can be taken to improve their performance. When they are used properly, mortality prediction models can save lives.

References

1. Teres D, Lemeshow S (1994) Why severity models should be used with caution. Crit Care Clin 10:93–110
2. Pollack MM, Patel KM, Ruttimann UE (1996) PRISM III: an updated paediatric risk of mortality score. Crit Care Med 24:743–752
3. Shann F, Pearson G, Slater A, Wilkinson K (1997) Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care. Intensive Care Med 23:201–207
4. Gemke R, Van Vught J (2002) Scoring systems in pediatric intensive care: PRISM III versus PIM. Intensive Care Med DOI 10.1007/s00134-001-1185-2
5. Hosmer DW, Lemeshow S (2000) Applied logistic regression. 2nd edn. Wiley, New York
6. Bertolini G, D’Amico R, Nardi D, Tinazzi A, Apolone G (2000) One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostatistics 5:251–253