Intensive Care Med (2005) 31:1229–1234
DOI 10.1007/s00134-005-2716-z

PEDIATRIC ORIGINAL

**Jolanda G. van Keulen**
**Reinoud J. B. J. Gemke**
**Kees H. Polderman**

# Effect of training and strict guidelines on the reliability of risk adjustment systems in paediatric intensive care

Received: 11 August 2004
Accepted: 13 June 2005
Published online: 6 July 2005


J. G. van Keulen · R. J. B. J. Gemke (✉)
Department of Paediatrics,
VU University Medical Center,
P.O. Box 7057, 1007 MB Amsterdam,
The Netherlands
e-mail: rjbj.gemke@vumc.nl
Tel.: +31-0-204443319
Fax: +31-0-204442918

K. H. Polderman
Department of Intensive Care,
VU University Medical Center,
P.O. Box 7057, 1007 MB Amsterdam,
The Netherlands

**Abstract** *Objective:* Many risk-adjustment systems have significant interobserver variability in everyday clinical practice. This can be partly corrected by strict guidelines and training. These issues have not been well studied in the paediatric setting. We assessed the reliability of two widely used paediatric scoring systems, the Paediatric Risk of Mortality (PRISM) and Paediatric Index of Mortality (PIM), before and after a special training program. *Design:* Prospective observational multi-centred cohort study. *Intervention:* Twenty-seven physicians from five paediatric intensive care units (PICUs) assessed severity of illness in 20 selected patients using PRISM and PIM scores before and after a special training program. Physicians were divided according to level of PICU experience: intensivists (>3 years experience, *n*=12), fellows (6–30 months experience, *n*=6) and residents (<6 months experience, *n*=9). Intraclass correlation was used to compare scoring reliability before and after training. *Measurements and results:* Wide variability in PRISM and PIM scoring was observed before training (intraclass correlation for PRISM scores 0.24–0.73, intraclass correlation for PIM scores 0.16–0.33). Training and implementation of guidelines led to significant increases in interobserver agreement (intraclass correlation 0.74–0.86 for PRISM and 0.88–0.95 for PIM scores), although some variability remained. *Conclusion:* Our results show that the reliability of PRISM and PIM risk adjustment systems in daily clinical practice is much lower than expected. Training and guidelines can significantly increase interobserver agreement. These factors should be taken into account when using these systems for benchmarking, or to compare quality of care between different PICUs.

**Keywords** Risk adjustment systems · Scoring systems · PRISM · PIM · Interobserver variability · Intraclass correlation

## Introduction

Risk adjustment systems, such as the Paediatric Risk of Mortality (PRISM) score and Paediatric Index of Mortality (PIM), are widely used in paediatric intensive care. These systems are used to allow assessment of severity of illness in heterogeneous patient groups in an objective manner, and to convert these risks into a numerical mortality risk. The purpose of their usage varies and may include comparison of severity of illness between different treatment arms in clinical trials as well as benchmarking, i.e. comparison of quality of care between different paediatric intensive care units (PICUs) using standardised mortality rates (i.e. mortality rates that have been adjusted for severity of illness). Both the PRISM and PIM scoring system have been developed and validated in tertiary PICUs. [1, 2, 3]. In some centres that were closely involved in developing these scoring systems, preliminary

**Table 1** Interobserver agreement of Paediatric Risk of Mortality (*PRISM*)- and Paediatric Index of Mortality (*PIM*)-score-based mortality risk (%) before and after implementation of guidelines and training program for all physicians ($n$=27)

| Before training | | | | | After training | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Patient no. | PRISM-score-based mortality risk[a] | | PIM-score-based mortality risk[b] | | Patient no. | PRISM-score-based mortality risk[a] | | PIM-score-based mortality risk[b] | |
| | Mean (SD) | Range | Mean (SD) | Range | | Mean (SD) | Range | Mean (SD) | Range |
| 1 | 0.6 (0.4) | 0.1–1.4 | 13.3 (25.1) | 2–46 | 1 | 73.0 (21.9) | 13–97 | 9.36 (3.57) | 4–20 |
| 2 | 1.8 (0.5) | 1.1–6.9 | 4.8 (3.1) | 0–19 | 2 | 22.1 (10.0) | 4–33 | 6.02 (1.16) | 1–8 |
| 3 | 4.7 (6.9) | 1.1–36.0 | 9.7 (8.8) | 5–27 | 3 | 1.45 (0.7) | 0–2 | 0.37 (0.32) | 0–1 |
| 4 | 2.7 (2.4) | 0.1–6.9 | 7.0 (9.1) | 1–51 | 4 | 7.73 (4.3) | 1–23 | 6.50 (1.49) | 2–10 |
| 5 | 3.2 (2.1) | 0.5–6.9 | 6.8 (4.5) | 2–16 | 5 | 22.7 (12.8) | 3–49 | 7.73 (9.97) | 4–57 |
| 6 | 0.6 (0.3) | 0.3–0.9 | 1.3 (1.3) | 1–5 | 6 | 17.4 (10.2) | 1–44 | 15.5 (6.47) | 6–32 |
| 7 | 2.5 (0.9) | 0.4–6.8 | 4.2 (1.2) | 1–7 | 7 | 90.8 (15.5) | 40–99 | 83.4 (15.2) | 38–98 |
| 8 | 17.2 (14.0) | 7.0–32 | 15.2 (15.3) | 2–56 | 8 | 55.1 (25.1) | 3–88 | 8.49 (4.48) | 1–17 |
| 9 | 0.5 (0.3) | 0.1–2.4 | 2.7 (1.8) | 1–5 | 9 | 28.6 (20.0) | 5–79 | 26.6 (16.6) | 7–75 |
| 10 | 33.0 (27.4) | 2.1–62.0 | 2.1 (2.0) | 1–9 | 10 | 8.46 (7.4) | 2–29 | 6.36 (3.10) | 3–19 |

data showed that the degree of interobserver reliability was acceptable [4, 5, 6]; however, these centres had a small and dedicated number of thoroughly trained professionals who were responsible for the scoring of patients. This form of organisation is likely to result in low interobserver reliability; however, the practical situation in numerous ICUs and PICUs throughout Europe is that severity scoring is performed by a varying number of residents, fellows, (paediatric) intensivists, paediatricians or nurses, with varying degrees of PICU experience and varying degrees of experience and training in the use of PRISM and PIM scores [7].

We previously demonstrated that significant degrees of interobserver variability in the use of PRISM and PIM scoring exist in everyday clinical practice, in physicians with different levels of training and experience [8]. Based on this observation we implemented a training program to improve the use of these risk adjustment scores.

This paper reports the results of this training program in improving accuracy of scoring and decreasing interobserver variability. All physicians who had participated in our first study received this training and were subsequently asked to participate in the present study.

## Methods

Physicians from six academic PICUs (tertiary referral centres) with residency and fellowship training programs were asked to participate in our study. Physicians were divided into three categories: residents ($n$=9) with limited experience in paediatric intensive care (average: 3 months; range: 6 weeks to 6 months); PICU fellows ($n$=6, average experience: range 6–30 months); and paediatric intensivists ($n$=12) with at least 3 years of full-time PICU experience. The charts of 20 patients that had been admitted to a single PICU in the course of a 1-year period were selected for scoring and randomly divided into two sets. The first set of ten charts was used before the training program. Charts of the second ten patients were used thereafter. The charts were selected to reflect typical PICU patients and were not chosen for difficulty of scoring. Relevant data from the medical charts and copies of blank data collection sheets from the PRISM and PIM scores were provided to all participating physicians. Subsequently, these physicians assessed the scores and filled out the data collection sheets. From the PRISM and PIM scores, calculated by the individual physicians, mean (SD) and range of scores were determined for each individual patient for the overall group of physicians and for each of the three different physician categories, according to methods described previously [9].

We observed significant interobserver variability in both PRISM and PIM scoring before implementation of our training program [8]. Based on these findings and on the specific problems in scoring interpretation that were identified, we implemented a training program and organised training sessions for all participating physicians. In these training sessions the guidelines of both scoring systems were extensively presented and discussed, and various pitfalls encountered in the first part of our study were discussed in detail. In addition, after the training sessions, the physicians received a summary of the guidelines for reference, as well as a summary of the subject matter of the training sessions.

Subsequently, the physicians were asked to assess the PRISM and PIM scores of the second series of 10 patients (group 2). Mean, standard deviation (SD), and range of PRISM and PIM for each patient were calculated for the whole group of physicians and according to level of experience and were compared with those before the training. Intraclass correlation was used to compare the reliability before and after implementation of training and strict guidelines for all physicians and per level of experience

Statistical assessment was performed using Student's $t$-test for unpaired variables for paired groups and by analysis of variance (ANOVA). Variability between observers was assessed by determining intraclass correlation coefficients.

Statistical significance was accepted for $p<0.05$. Excel (Microsoft, Redman, Wash.) and SPSS 9.0 (SPSS, Chicago, Ill.) software was used for all calculations.

## Results

The results for the whole group of physicians before and after training are shown in Tables 1 and 2. The PRISM- and PIM-based mortality risks in individual patients before and after training and implementation of guidelines are shown in Table 1. The intraclass correlation for

**Table 2** Intraclass correlation for PRISM- and PIM-score-based mortality risk before and after administration of guidelines and training

| | Before training and guidelines | | | | After training and guidelines | | | |
|---|---|---|---|---|---|---|---|---|
| | PRISM | | PIM | | PRISM | | PIM | |
| | ICC | 95% CI | ICC | 95% CI | ICC | 95% CI | ICC | 95% CI |
| All physicians (*n*=27) | 0.51 | 0.32–0.78 | 0.18 | 0.08–0.46 | 0.80** | 0.65–0.93 | 0.89** | 0.80–0.97 |
| Residents (*n*=9) | 0.24* | 0.07–0.57 | 0.33 | 0.14–0.65 | 0.77** | 0.59–0.94 | 0.88** | 0.77–0.96 |
| Fellows (*n*=6) | 0.40 | 0.16–0.73 | 0.33 | 0.11–0.67 | 0.74** | 0.51–0.91 | 0.95** | 0.89–0.98 |
| Intensivists (*n*=12) | 0.73 | 0.57–0.91 | 0.16 | 0.11–0.39 | 0.86** | 0.73–0.95 | 0.88** | 0.76–0.96 |

*$p<0.01$ in comparison with intensivists
**$p< 0.05$ in comparison before and after training

**Table 3** Interobserver agreement of PRISM-score-based mortality risk (%) *after* guidelines and training divided according to different levels of experience

| Patient no. | All physicians (*n*=27) | | Residents (*n*=9) | | | Fellows (*n*=6) | | | Intensivists (*n*=12) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean (SD) | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| 1 | 73.0 (21.9) | 13–97 | 70.8 | 16.8 | 44–88 | 61.8 | 29.0 | 13–90 | 80.3 | 20.2 | 25–97 |
| 2 | 22.1 (9.97) | 4–33 | 22.6 | 11.2 | 7–33 | 19.4 | 13.1 | 4–33 | 23.0 | 7.83 | 10–33 |
| 3 | 1.45 (0.67) | 0–2 | 1.62 | 0.67 | 0–2 | 1.14 | 0.65 | 0–2 | 1.47 | 0.68 | 0–2 |
| 4 | 7.73 (4.25) | 1–23 | 8.93 | 6.15 | 2–23 | 5.69 | 2.14 | 3–8 | 7.85 | 3.11 | 1–12 |
| 5 | 22.7 (12.8) | 3–49 | 21.8 | 9.52 | 3–34 | 15.2 | 6.91 | 6–22 | 27.1 | 15.6 | 5–49 |
| 6 | 17.4 (10.2) | 1–44 | 21.9 | 12.6 | 3–44 | 12.7 | 6.21 | 4–19 | 16.2 | 9.18 | 1–39 |
| 7 | 90.8 (15.5) | 40–99 | 89.6 | 17.3 | 51–98 | 83.0 | 22.0 | 40–97 | 95.6 | 8.43 | 70–99 |
| 8 | 55.1 (25.1) | 3–88 | 50.6 | 30.5 | 3–88 | 38.8 | 22.9 | 8–72 | 66.6 | 16.2 | 17–76 |
| 9 | 28.6 (20.0) | 5–79 | 28.5 | 18.4 | 18–76 | 23.1 | 18.5 | 5–58 | 31.4 | 22.7 | 5–80 |
| 10 | 8.46 (7.36) | 2–29 | 8.49 | 6.71 | 3–25 | 4.29 | 3.25 | 1–10 | 10.5 | 8.73 | 2–29 |

**Table 4** Interobserver agreement of PIM-score-based mortality risk (%) *after* guidelines and training divided according to different levels of experience

| Patient no. | All physicians (*n*=27) | | Residents (*n*=9) | | | Fellows (*n*=6) | | | Intensivists (*n*=12) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean (SD) | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range |
| 1 | 9.36 (3.57) | 4–20 | 9.63 | 3.06 | 4–13 | 9.45 | 4.49 | 7–18 | 9.11 | 3.74 | 6–20 |
| 2 | 6.02 (1.16) | 1–8 | 6.18 | 0.80 | 5–7 | 5.64 | 0.83 | 5–6 | 6.08 | 1.51 | 1–8 |
| 3 | 0.37 (0.32) | 0–1 | 0.22 | 0.04 | 0–1 | 0.46 | 0.42 | 0–1 | 0.43 | 0.37 | 0–1 |
| 4 | 6.50 (1.49) | 2–10 | 6.04 | 1.59 | 3–7 | 6.83 | 0.21 | 6–7 | 6.67 | 1.77 | 2–10 |
| 5 | 7.73 (9.97) | 4–57 | 11.2 | 17.4 | 4–57 | 5.83 | 0.31 | 5–6 | 6.13 | 0.67 | 5–7 |
| 6 | 15.5 (6.47) | 6–32 | 19.9 | 8.80 | 13–32 | 12.2 | 0.71 | 11–13 | 13.8 | 4.18 | 6–22 |
| 7 | 83.4 (15.2) | 38–98 | 85.7 | 13.0 | 65–98 | 78.2 | 15.0 | 65–92 | 84.3 | 17.4 | 38–97 |
| 8 | 8.49 (4.48) | 1–17 | 10.7 | 4.48 | 4–17 | 6.01 | 2.90 | 1–8 | 8.07 | 4.62 | 1–16 |
| 9 | 26.6 (16.6) | 7–75 | 26.8 | 13.5 | 11–57 | 21.3 | 1.41 | 19–22 | 29.1 | 22.3 | 7–75 |
| 10 | 6.36 (3.10) | 3–19 | 8.46 | 3.99 | 6–19 | 5.13 | 1.77 | 3–7 | 5.40 | 2.05 | 3–8 |

PRISM and PIM scoring before and after training are shown in Table 2. The PRISM- and PIM-based mortality risks divided by category of physicians are shown in Table 3 (PRISM scores) and Table 4 (PIM scores).

As can be seen in the crude data in Table 1 and the calculations in Table 2, we observed substantial interobserver variability in both PRISM and PIM scoring before implementation of our training program. For PRISM scores average intraclass correlation was 0.51 (range 0.32–0.78); for PIM scores the average intraclass correlation was only 0.18 (range 0.08–0.46). This variability occurred in both experienced and inexperienced physicians [7].

Interobserver agreement for both PRISM- and PIM-score-based risk assessment improved significantly after implementation of guidelines and training. The intraclass correlation after training varied from 0.74 to 0.86 for the PRISM scores, and 0.88 to 0.95 for the PIM score. The increase in intraclass correlation following training was statistically significant ($p<0.01$).

When subdivided according to levels of PICU experience (consultants, fellows and residents, respectively) we found that before training the intra-class correlations for PRISM scoring were significantly lower for residents than in the group of intensivists ($p<0.01$; Table 2). No such differences were observed for the PIM score (indeed

residents appeared to perform slightly though not significantly better). Compared with the measurements before training, there was a substantial decrease in interobserver variability in all three categories of physicians, as indicated by the significant difference in intraclass correlations for the whole group of physicians ($p<0.01$) and per level of experience ($p<0.05$ for intensivists, fellows and residents, respectively).

The results for individual patients divided by category of physicians are shown in Table 3 (PRISM scores) and Table 4 (PIM scores). Following training, the differences in performance between the three groups decreased, with intraclass correlation $\geq 0.74$ observed in all groups of physicians.

## Discussion

The results of our study demonstrate that training and implementation of strict guidelines are required for reliable assessment of the PRISM and PIM scores. Those physicians using these risk adjustment systems for PICU quality assessment and benchmarking should take this into account.

Our assessment of variability before training revealed a surprisingly high level of variability, with an average intraclass correlation of 0.51 (range 0.32–0.78) for the PRISM score and an average intraclass correlation of only 0.18 (range 0.08–0.46) for the PIM score. These figures were well below our expectations; however, they are likely to reflect the reality in numerous European PICUs where regular training in use of these risk adjustment systems has not been rigorously implemented. Moreover, physicians with varying degrees of experience from different medical centres participated in our study, which increases the likelihood that our results reflect the actual situation.

Our second measurement showed much improved results. The average intraclass correlation after training and guidelines was 0.80 (range 0.65–0.93) for the PRISM score and 0.89 (range 0.80–0.97) for the PIM score. The changes tended to be more prominent in less experienced physicians but were also observed in paediatric intensivists who have at least >3 years of PICU experience. Both series of patients were randomly selected to represent the spectrum of patients that are admitted to the PICU, so it is unlikely that a more complicated group that might incur a lower intraclass correlation may have been selected for the first measurement. Indeed, the average scores were actually somewhat higher in patients that were selected after training, which is likely to increase the likelihood of error. This implies that the effects of training are likely to have been somewhat underestimated in our study.

However, even after training and implementation of guidelines, a significant degree of variability in scoring persisted, even in experienced intensivists with comparable training, experience and background; therefore, it seems likely that some degree of variability is inherent in PRISM and PIM scoring, at least in current clinical practice PICUs in the Netherlands. There are no important differences in the way in which these systems are used between PICUs in the Netherlands and most other European PICUs; therefore, our results are likely to reflect the situation in PICUs with similar forms of organisation throughout Europe.

An interesting observation is the difference in variability and intraclass correlations between PRISM and PIM scores. Before training, intraclass correlation was lower for PIM compared with PRISM, whereas after training PIM had a slightly higher intraclass correlation. The reason for these differences are unclear. In theory, the observation that intraclass correlation was initially lower for the PIM score may be explained by the fact that the PRISM scores were the first to be implemented in everyday clinical practice, and therefore have been used for longer periods of time. Their earlier introduction and the period during which PRISM scores were the only risk adjustment system available for the paediatric population may have made PRISM scores somewhat more familiar to paediatric clinicians, even though PIM scores have also been used for several years. Our observation that especially experienced intensivists had comparatively high intraclass correlations for PRISM scoring, with far lower scores for PIM scoring, lends some credence to this hypothesis. An additional factor could be the lower number of variables in the PIM score, which could have led to a greater proportional effect of individual errors. This could also help explain the greater improvement in PIM scoring associated with training: if the lower number of variables in PIM scoring led to greater proportional disagreements before training, any reductions in these errors after training would also lead to greater proportional improvements in intraclass correlation; however, these potential explanations remain speculative, as direct comparisons between PIM and PRISM scoring were not made, and reasons for potential differences were not determined in our study.

Previous studies comparing the reliability of PIM and PRISM scores have reported that both are adequate indicators of probability of mortality for heterogeneous paediatric patient groups [10, 11, 12], with the PIM score performing perhaps marginally better in paediatric cardiac surgery patients [12]. In recent years PICU mortality has decreased, and overall outcome in paediatric critical care has improved significantly. Long-term outcome has also improved, with good functional recovery and quality of life for surviving patients [13]. This has led to a relative overestimation of mortality by both PIM and PRISM risk adjustment systems. The PIM score has recently been revised to take improvements in outcome into account [14].

A potential limitation of our findings is that there was no attempt to determine overall "accuracy" by comparison with a gold standard, i.e. if all observers would make the same mistake, overall agreement would be good, whereas accuracy would be poor. This again might have led to underestimation of the problems with severity scoring; however, to address this issue would require a separate study in which scores are compared with a gold standard, which could consist of a panel of experts (who would have to score all patients according to pre-defined criteria and agree on all issues).

Another potential limitation is that different patients had to be used for the two measurements of variability, to prevent physicians "remembering" issues about individual patients which would have influenced the results. In theory, one of the groups of patients could have been more "difficult" to score, leading to greater variability. Indeed, average scores were slightly higher for the second measurement, indicating the presence of a number of more severely ill patients. In theory, this could imply that variability after training may have been somewhat overestimated in our study; however, the fact that significant variability occurred also in patients with lower scores during the second measurement, and that variability as a percentage of the score in each patient was fairly constant, makes it highly unlikely that this would have significantly affected our overall results and conclusions.

Reliability of risk adjustment may be improved, and variability decreased, if severity of illness scoring is performed by a restricted number of dedicated individuals who are well trained and regularly audited; however, the efficacy of this strategy needs to be determined in future studies, and does not reflect the current overall situation in European PICUs. Our present study shows that substantial improvements in reliability may well be obtained using a rigorous but relatively uncomplicated training program and guidelines. Continued reliability may well require regular updates and audits.

The observations in this multi-centre study are in accordance of previous observations by our group and others [9, 15, 16, 17, 18] on everyday use of the APACHE II scoring system, which is the most widely used risk adjustment system in adult ICUs. The use of this scoring system is associated with interobserver variability of up to 30% in everyday clinical practice [16, 17]. This decreases to around 15% after implementation of guidelines and training [9].

Previous authors have suggested that an ICC above 0.80 should be considered acceptable in a clinical setting [19, 20]. In our study neither the PIM nor the PRISM score reached this value before training. After training both scores realised intraclass correlations $\geq$0.80 (albeit only just in the case of the PRISM score). Nevertheless, a degree of variability remained even after training, a fact that physicians using these scores should be aware of even if the degree of variability is deemed acceptable. Some authors have suggested that risk adjustment systems could be used to predict outcome in individual patients [21], although their use for this purpose remains controversial both in the adult and paediatric populations [22, 23]. If attempts are made to predict risk of death in individual patients, issues of reliability of assessment and interobserver variability become even more important.

Another novel application of severity scoring systems is for selecting patients who might gain the greatest benefits from specific treatments. An example of this is the use of APACHE II scores to select patients with severe sepsis for treatment with activated protein C [24]. The use of APACHE II scores in this way is based on observations from the PROWESS trial [23]. This study, which reported a significant decrease in mortality in a large group of patients with severe sepsis treated with activated protein C, observed greater benefits in patients with higher APACHE II scores compared with the overall group [25]; however, the use of risk adjustment systems for such purposes has been challenged on various grounds [26]. Systems such as APACHE II, PRISM and PIM were designed for outcome prediction in large groups of patients, and have never been validated for risk assessment in individual patients. In our opinion, great caution should be taken when making decisions on allocation of resources and treatments in individual patients based on risk adjustment systems. This view is reinforced by observations that organisational changes, case mix of patients and the transfer of patients between units can substantially affect various benchmarking tools to assess ICU performance, including the frequently used standardised mortality ratio [27, 28].

## Conclusion

In conclusion, although PRISM and PIM scores are valuable tools in paediatric intensive care, it is important to realize that reliability of risk adjustment systems in everyday clinical practice is highly dependent on the implementation of training, guidelines and regular audit of these scoring systems. Even when these precautions are taken, a degree of interobserver and even intraobserver variability is likely to persist. The observations in adult ICUs and our current findings in the paediatric setting underscores the importance of being aware of the limitations of risk adjustment systems, especially when the are used for benchmarking and to assess quality of care in the (paediatric) ICU.

**Acknowledgement** The authors thank all participating physicians form de Dutch PICUs (Academic Medical Centre Amsterdam, University Medical Centre, Utrecht, Leiden University Medical Centre, Academic Hospital Groningen, Academic Hospital Maastricht, VU University Medical Centre) and the University Hospital Leuven, Belgium.

## References


1. Pollack MM, Ruttimann UE, Getson PR (1988) Pediatric risk of mortality score (PRISM) score. Crit Care Med 16:1110–1116
2. Shann F, Pearson G, Slater A, Wilkinson K (1997) Pediatric index of mortality (PIM): a mortality prediction model for children in intensive care. Intensive Care Med 23:201–207
3. Pollack MM, Patel KM, Ruttimann UE (1997) The pediatric risk of mortality III: Acute Physiology Score (PRISM III-APS): a method of assessing physiologic instability for pediatric intensive care unit patients. J Pediatr 131:575–581
4. Gemke RJBJ, Bonsel GJ (1995) The Pediatric Intensive Care Assessment of Outcome (PICASSO) study group. Comparative assessment of pediatric intensive care: a national multicenter study. Crit Care Med 23:238–245
5. Slater A, Shann F, Gearson G (2003) PIM2: a revised version of the Pediatric Index of Mortality. Intensive Care Med 29:278–285
6. Tibby SM, Taylor D, Festa M, Hanna S, Hatherill M, Jones G, Habibi P, Durward A, Murdoch IA (2002) A comparison of three scoring systems for mortality risk among retrieved intensive care patients. Arch Dis Child 87:421–425
7. Gemke RJBJ, Bonsel GJ, van Vught AJ (1994) Effectiveness and efficiency of a Dutch pediatric intensive care unit: validity and application of the Pediatric Risk of Mortality score. Crit Care Med 22:1477–1484
8. van Keulen JG, Polderman KH, Gemke RJBJ (2005) Reliability of PRISM and PIM scores in paediatric intensive care. Arch Dis Child 90:211–214
9. Polderman KH, Jorna EMF, Girbes ARJ (2001) Inter-observer variability in APACHE II scoring: effect of strict guidelines and training. Intensive Care Med 27:1365–1369
10. Gemke RJ, van Vught J (2002) Scoring systems in pediatric intensive care: PRISM III versus PIM. Intensive Care Med 28:204–207
11. Shann F (2002) Are we doing a good job: PRISM, PIM and all that. Intensive Care Med 28:105–107
12. Jones GD, Thorburn K, Tigg A, Murdoch IA (2000) Preliminary data: PIM vs PRISM in infants and children post cardiac surgery in a UK PICU. Intensive Care Med 26:145 (Letter)
13. Taylor A, Butt W, Ciardulli M (2003) The functional outcome and quality of life of children after admission to an intensive care unit. Intensive Care Med 29:795–800
14. Slater A, Shann F, Pearson G (2003) PIM2: a revised version of the Paediatric Index of Mortality. Intensive Care Med 29:278–285
15. Polderman KH, Thijs LG, Girbes ARJ (1999) Interobserver variability in the use of APACHE II scores. Lancet 353:380
16. Polderman KH, Christiaans HM, Wester JP, Spijkstra JJ, Girbes AR (2001) Intraobserver variability in APACHE II scoring. Intensive Care Med 27:1550–1552
17. Polderman KH, Girbes ARJ, Thijs LG, Strack van Schijndel RJM (2001) Accuracy and reliability of APACHE II scoring in two intensive care units. Problems and pitfalls in the use of APACHE II and suggestions for improvement. Anaesthesia 56:47–50
18. Chen LM, Martin CM, Morrison TL, Sibbald WJ (1999) Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. Crit Care Med 27:1999–2004
19. Fleiss JL, Shrout PE (1977) The effects of measurement errors on some multivariate procedures. Am J Public Health 67:1188–1191
20. Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London, pp 277–306
21. Lemeshow S, Le Gall JR (1994) Modeling the severity of illness of ICU patients. J Am Med Assoc 272:1049–1055
22. Teres D, Lemeshow S (1994) Why severity models should be used with caution. Crit Care Clin 10:93–110
23. Randolph AG (1997) Paediatric index of mortality (PIM): Do we need another paediatric mortality prediction score? Intensive Care Med 23:141–142
24. Manns BJ, Lee H, Doig CJ, Johnson D, Donaldson C (2002) An economic evaluation of activated protein C treatment for severe sepsis. N Engl J Med 347:993–1000
25. Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, Steingrub JS, Garber GE, Helterbrand JD, Ely EW, Fisher CJ Jr, Recombinant Human Protein C Worldwide Evaluation in Severe Sepsis (PROWESS) Study Group (2001) Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med 344:699–709
26. Polderman KH, Girbes ARJ (2004) Drug intervention trials in sepsis: divergent results. Lancet 363:1721–1723
27. van Zanten AR, Polderman KH (2004) Organizational changes in a single intensive care unit affect benchmarking. Ann Intern Med 140:674–675
28. Rosenberg AL, Hofer TP, Strachan C, Watts CM, Hayward RA (2003) Accepting critically ill transfer patients: adverse effect on a referral center's outcome and benchmark measures. Ann Intern Med 138:882–890