Research

Open Access

Can generic paediatric mortality scores calculated 4 hours after admission be used as inclusion criteria for clinical trials?

Stéphane Leteurtre[1], Francis Leclerc[2], Jessica Wirth[3], Odile Noizet[4], Eric Magnenant[4], Ahmed Sadik[5], Catherine Fourier[5] and Robin Cremer[5]

[1]Paediatric Intensivist, Paediatric Intensive Care Unit, University Hospital of Lille, and SAMU, Lille, France
[2]Professor, Director, Paediatric Intensive Care Unit, University Hospital of Lille, Lille, France
[3]Resident, Paediatric Intensive Care Unit, University Hospital of Lille, Lille, France
[4]Clinical Fellow, Paediatric Intensive Care Unit, University Hospital of Lille, Lille, France
[5]Paediatric Intensivist, Paediatric Intensive Care Unit, University Hospital of Lille, Lille, France

Corresponding author: Francis Leclerc, fleclerc@chru-lille.fr

Received: 07 November 2003

Revisions requested: 16 January 2004

Revisions received: 07 April 2004

Accepted: 20 April 2004

Published: 21 May 2004

Critical Care 2004, **8**:R185-R193 (DOI 10.1186/cc2869)

This article is online at: http://ccforum.com/content/8/4/R185

Abstract

Introduction Two generic paediatric mortality scoring systems have been validated in the paediatric intensive care unit (PICU). Paediatric RISk of Mortality (PRISM) requires an observation period of 24 hours, and PRISM III measures severity at two time points (at 12 hours and 24 hours) after admission, which represents a limitation for clinical trials that require earlier inclusion. The Paediatric Index of Mortality (PIM) is calculated 1 hour after admission but does not take into account the stabilization period following admission. To avoid these limitations, we chose to conduct assessments 4 hours after PICU admission. The aim of the present study was to validate PRISM, PRISM III and PIM at the time points for which they were developed, and to compare their accuracy in predicting mortality at those times with their accuracy at 4 hours.

Methods All children admitted from June 1998 to May 2000 in one tertiary PICU were prospectively included. Data were collected to generate scores and predictions using PRISM, PRISM III and PIM.

Results There were 802 consecutive admissions with 80 deaths. For the time points for which the scores were developed, observed and predicted mortality rates were significantly different for the three scores ($P < 0.01$) whereas all exhibited good discrimination (area under the receiver operating characteristic curve ≥0.83). At 4 hours after admission only the PIM had good calibration ($P = 0.44$), but all three scores exhibited good discrimination (area under the receiver operating characteristic curve ≥0.82).

Conclusions Among the three scores calculated at 4 hours after admission, all had good discriminatory capacity but only the PIM score was well calibrated. Further studies are required before the PIM score at 4 hours can be used as an inclusion criterion in clinical trials.

Keywords: intensive care, mortality, prediction model

Introduction

Adjustment to severity is considered important in clinical trials for ensuring comparability between groups. Generic mortality scoring systems for children admitted to intensive care units (ICUs) have been developed for use at specific time points in the ICU stay. Two systems have been validated in paediatric ICUs (PICUs): the Paediatric RISk of Mortality (PRISM) and the Paediatric Index of Mortality (PIM). The PRISM, which is used in PICUs worldwide, requires an observation period of 24 hours [1], and the updated PRISM III score [2] measures

AUC = area under the receiver operating characteristic curve; df = degrees of freedom; ICU = intensive care unit; PICU = paediatric intensive care unit; PIM = Paediatric Index of Mortality; PRISM = Paediatric RISk of Mortality; SMR = standardized mortality ratio.

severity at two time points (12 and 24 hours) during the PICU stay. The PIM and the recently updated PIM2 scores are calculated 1 hour after admission [3,4]. The 12–24 hour period of observation has been a criticism levelled at the PRISM scoring system, and it has been speculated that it may diagnose rather than predict death [4,5]. With the PIM and PIM2 scores, the single measurement of values shortly after admission is susceptible to random variation [6] or may reflect a transient state resulting from interventions during transport [7].

Severity models have been used for time periods different from those for which the scores were developed [8]. In children with meningococcal septic shock, Castellanos-Ortega and coworkers [9] recorded the worst values for each variable included in the Glasgow Meningococcal Septicaemia Prognostic Score, the Malley score, and the PIM score over the first 2 hours in the PICU. Indeed, early identification of patients who could benefit from therapeutic interventions may be useful [9].

We hypothesized that an intermediate observation period (we arbitrarily chose a time point of 4 hours after PICU admission) would be a good compromise between two objectives – to take into account a short period of stabilization after a patient's admission to the PICU and to obtain an accurate measure of illness severity in the PICU. To our knowledge, no study has ever evaluated the accuracy of generic paediatric scoring systems in predicting death for the whole PICU population, and for time periods different from those for which the scores were developed.

The aim of the present study was to externally validate the PRISM, PRISM III and PIM scores at their intended time points, and to compare their accuracy in predicting mortality at those times with their accuracy at a different time period, namely 4 hours after admission.

Methods

All consecutive patients admitted to our university hospital PICU from June 1998 through to May 2000 were included unless they met the following exclusion criteria: admission in a state requiring cardiopulmonary resuscitation without achieving stable vital signs for at least 2 hours; admission for scheduled procedures normally done in other hospital wards; prematurity; and age more than 18 years.

Standard documentation and training were provided to all PICU medical staff. Data were prospectively collected to generate scores and predictions for the time periods for which the scores were developed (i.e. PIM at 1 hour, PRISM at 24 hours, PRISM III at 12 hours, and PRISM III at 24 hours) and to generate scores and predictions for a different time point (i.e. 4 hours after admission) [1,2,4]. The PIM2 score was not evaluated because it had not yet been reported when we began the study. The outcome measure was death or survival at discharge from the PICU. The probabilities of death were calculated at different time points (Table 1). To generate a prediction for the PRISM III 4-hour score, we used the PRISM III 12-hour equation (1996 version). In order to compare observed with expected mortality and to estimate the calibration of the scores, a Hosmer–Lemeshow goodness-of-fit test with five degrees of freedom (df; we considered five classes of mortality probability: 0% to <1%, 1% to <5%, 5% to <15%, 15% to <30%, and ≥30%) was performed [1]. According to this test, the *P* value is greater than 0.05 if the model is well calibrated; the greater the *P* value, the better the model fits [10].

Table 1

Time of scoring

Score	1 hour	4 hours	12 hours	24 hours
PIM	x	x		
PRISM		x		x
PRISM III		x	x	x

PIM, Paediatric Index of Mortality [4] ; PRISM, Paediatric RISk of Mortality [1,2].

The areas under the receiver operating characteristic curve (AUCs) and their standard errors were calculated to estimate the discrimination of the scores. An AUC ≥0.7 is generally considered acceptable, ≥0.8 as good, and ≥0.9 as excellent [11,12]. Standardized mortality ratios (SMRs) and their comparison to 1 were calculated [13]. To study the effect of length of stay on calibration and discrimination of the three scores, calibration was calculated each day and discrimination at days 5, 10 and 20 after admission. For a patient who had died on day_x, the PICU outcome-dependent variable was considered as survival at day_{x-i} (i = 1, 2, 3...).

Statistical analyses were performed using the Statistical Program for Social Science (SPSS Inc., Chicago, IL, USA).

Results

There were 802 consecutive admissions with 80 deaths (10%). Medical patients represented 81%. The median age was 26 months (interquartile range 8–92 months) and the median length of stay was 2 days (interquartile range 1–6 days). The frequencies of the additional variables in the PRISM III score were as follows: nonoperative cardiovascular disease 4.2%, cancer 5.9%, chromosomal anomalies 2%, previous PICU admission 6.5%, pre-PICU cardiopulmonary resuscitation 4.2%, postoperative surgical procedure 19%, acute diabetes 0.7%, and admission from routine care area 13%.

For the time periods for which the scores were developed, the three scores had poor calibration ($P < 0.01$ for each), with large differences between the χ^2 goodness-of-fit test values (Table 2). We observed underestimations of mortality in the

Table 2

Hosmer–Lemeshow goodness-of-fit test values and AUCs: time point for which the scores were developed

Score	Hour	χ^2 goodness-of-fit test value	P (df = 5)	AUC	Standard error
PIM	1	18.7	0.002	0.83	0.03
PRISM	24	15.9	0.007	0.86	0.02
PRISM III	12	27.1	$<10^{-4}$	0.91	0.02
PRISM III	24	33.0	$<10^{-5}$	0.92	0.02

AUC, area under the receiver operating characteristic curve; df, degrees of freedom; PIM, Paediatric Index of Mortality [4]; PRISM, Paediatric RISk of Mortality [1,2].

low mortality risk groups (risk 1% to <5% and risk 5% to <15%), and an overestimation in the group with very high risk for mortality (risk ≥30%). SMRs varied from 1.03 to 1.39, but only the SMR for the PIM 1-hour assessment was significantly greater than 1 (Table 3). All scores exhibited good discrimination (Table 2).

At 4 hours PIM had good calibration ($P = 0.44$). Conversely, both PRISM and PRISM III had poor calibration at 4 hours ($P < 0.01$), with significant differences between observed and predicted mortality (Tables 4 and 5). Expected mortality with PRISM and PRISM III underestimated the observed mortality in the groups at low risk for mortality. SMRs varied from 1.17 to 1.57, and were significantly greater than 1 except for the PIM 4-hour assessment (Table 5). All scores exhibited good discrimination (Table 4).

For the time points for which the scores were developed, study of the length of stay showed good calibration for the PIM 1-hour assessment between days 3 and 28, for the PRISM 24-hour assessment between days 11 and 22, for the PRISM III 12-hour assessment between days 51 and 58, and for the PRISM III 24-hour assessment between days 10 and 11 (Fig. 1a). For the different time point examined (i.e. 4 hours), study of the length of stay showed good calibration for PIM from day 4 until discharge, for PRISM between days 2 and 15, and for PRISM III between days 3 and 10 (Fig. 1b). For both time periods, study of the length of stay showed that the AUC for all scores, both for the time points for which the scores were developed (Fig. 2a) and at 4 hours (Fig. 2b), exceeded 0.80.

With regard to the poor calibration identified in some of the assessments, retrospective analysis of patients who died was performed for the classes of mortality probability for which the χ^2 value exceeded 2.5. A χ^2 value of 11 was needed to obtain statistical calibration with the five classes of mortality probability. For these deceased patients we analyzed length of stay and comorbidities (cancer, prematurity, and chronic cardiac, respiratory, neurological and digestive diseases). Chronic organ disease was defined as disease with or without organ failure, requiring multiple admissions (to paediatric department or day care center) and requiring supervision by a subspecialist in paediatrics. A χ^2 value over 2.5, which indicates a significant difference between observed and predicted probability of death in a mortality class, was observed for 55 deceased patients. In this subpopulation, the median length of stay was significantly different from that in the other 25 deceased patients (7 days versus 1 day, respectively; $P < 0.001$), and only seven (13%) had a pre-ICU cardiac massage versus 18 (72%) in the other deceased patients ($P < 0.000001$). In these 55 patients, only 6–11% of the above mentioned comorbidities were taken into account in the probability of death calculated with the different scores.

Discussion

In this single unit study, discrimination of the PIM, PRISM and PRISM III scores was good whereas calibration was poor for the time points for which the scores were developed. At 4 hours, only the PIM score had good discrimination and calibration.

Both discrimination and calibration must be considered when evaluating the performance of scoring systems [14].

Discrimination measures the predictive performance of scoring systems, and when the outcome is dichotomous it is usually described by a receiver operating characteristic curve. In the studies that compared the original PIM, PRISM and PRISM III scores, the AUCs were as follows: ≥0.7 for the PIM and PRISM III scores [15]; ≥0.8 for the PIM score, and ≥0.9 for the PRISM and PRISM III scores [16]; and between 0.83 and 0.87 for the pre-ICU PRISM, PIM and PRISM scores [5]. Those findings are similar to ours. However, for Zhu and coworkers [17] AUC was not as sensitive to differences in ICU care as the Hosmer–Lemeshow goodness-of-fit test.

Gemke and van Vught [15] externally validated the PRISM III and PIM scores ($n = 303$ patients, 20 deaths); the goodness-of-fit test values with 10 deciles of mortality risk were 10.8 ($P = 0.21$, df = 8) for the PRISM III 12-hour assessment, 13.3 ($P = 0.10$ [not $P = 0.21$, as was published], df = 8) for the PRISM III 24-hour assessment, and 4.92 ($P = 0.77$, df = 8) for the PIM score. However, the P values that we calculated from these data using the five conventional mortality risk categories were

Table 3

Hosmer–Lemeshow goodness-of-fit test values: time point for which the scores were developed

Score	Mortality probability classes (%)	Expected deaths	Observed deaths	Observed survivors	Expected survivors	χ^2 goodness-of-fit test values	SMR (95% CI)
PIM 1 hour	<1	1.3	3	175	176.7	2.2	
	1 to <5	8.3	15	357	363.7	5.5	
	5 to <15	12.2	23	127	137.9	10.4	
	15 to <30	13.0	15	47	49.0	0.4	
	≥30	22.8	24	16	17.2	0.1	
Totals		**57.6**	**80**	**722**	**744.7**	**18.7**	**1.39 (1.10–1.73)***
PRISM 24 hours	<1	1.5	4	237	239.5	4.2	
	1 to <5	7.7	16	313	321.3	9.2	
	5 to <15	10.2	9	112	110.8	0.2	
	15 to <30	8.5	7	33	31.5	0.3	
	≥30	49.5	44	27	21.5	2.0	
Totals		**77.4**	**80**	**722**	**724.6**	**15.9**	**1.03 (0.82–1.29)**
PRISM III 12 hours	<1	1.8	4	422	424.2	2.7	
	1 to <5	4.4	14	175	184.6	21.4	
	5 to <15	7.5	10	75	77.5	0.9	
	15 to <30	7.2	5	27	24.8	0.9	
	≥30	51.1	47	23	18.9	1.2	
Totals		**72.0**	**80**	**722**	**730.0**	**27.1**	**1.11 (0.88–1.38)**
PRISM III 24 hours	<1	1.7	3	474	475.3	1.0	
	1 to <5	3.3	12	133	141.7	23.5	
	5 to <15	7.0	14	72	79.0	7.7	
	15 to <30	6.2	5	25	23.8	0.3	
	≥30	48.6	46	18	15.4	0.6	
Totals		**66.8**	**80**	**722**	**735.2**	**33.0**	**1.20 (0.95–1.49)**

*Significantly greater than 1 ($P = 0.002$) [13]. CI, confidence interval; PIM, Paediatric Index of Mortality [4]; PRISM, Paediatric RISk of Mortality [1,2]; SMR, standardized mortality ratio.

Table 4

Hosmer–Lemeshow goodness-of-fit test values and AUCs: 4 hours

Score	Hour	χ^2 goodness-of-fit test value	P (df = 5)	AUC	Standard error
PIM	4	4.8	0.44	0.86	0.02
PRISM	4	47.3	$<10^{-8}$	0.82	0.03
PRISM III	4	47.2	$<10^{-9}$	0.90	0.02

AUC, area under the receiver operating characteristic curve; df, degrees of freedom; PIM, Paediatric Index of Mortality [4]; PRISM, Paediatric RISk of Mortality [1,2].

Table 5

Hosmer-Lemeshow goodness of fit test values: 4 hours

Score	Mortality probability classes (%)	Expected deaths	Observed deaths	Observed survivors	Expected survivors	χ^2 goodness-of-fit test values	SMR (95% CI)
PIM 4 hours	<1	0.9	1	140	140.1	0.0	
	1 to <5	7.9	13	368	373.1	3.4	
	5 to <15	13.1	14	144	144.9	0.1	
	15 to <30	14.4	18	50	53.6	1.1	
	≥30	32.2	34	20	21.8	0.2	
Total		**68.5**	**80**	**722**	**733.5**	**4.8**	**1.17 (0.93–1.45)**
PRISM 4 hours	<1	1.9	8	299	305.1	19.7	
	1 to <5	7.7	20	321	333.3	20.1	
	5 to <15	6.6	12	68	73.4	4.8	
	15 to <30	6.5	10	21	24.5	2.4	
	≥30	28.3	30	13	14.7	0.3	
Totals		**51.0**	**80**	**722**	**751.0**	**47.3**	**1.57 (1.24–1.95)***
PRISM III 4 hours	<1	1.9	6	455	459.1	8.9	
	1 to <5	4.3	16	167	178.7	32.6	
	5 to <15	6.3	12	60	65.7	5.7	
	15 to <30	7.0	8	25	26.0	0.2	
	≥30	37.2	38	15	15.8	0.1	
Totals		**56.7**	**80**	**722**	**745.3**	**47.2**	**1.41 (1.12–1.76)***

*Significantly greater than 1 ($P < 0.0001$ for PRISM at 4 hours and $P = 0.0025$ for PRISM III at 4 hours) [13]. CI, confidence interval; PIM, Paediatric Index of Mortality [4]; PRISM, Paediatric RISk of Mortality [1,2]; SMR, standardized mortality ratio.

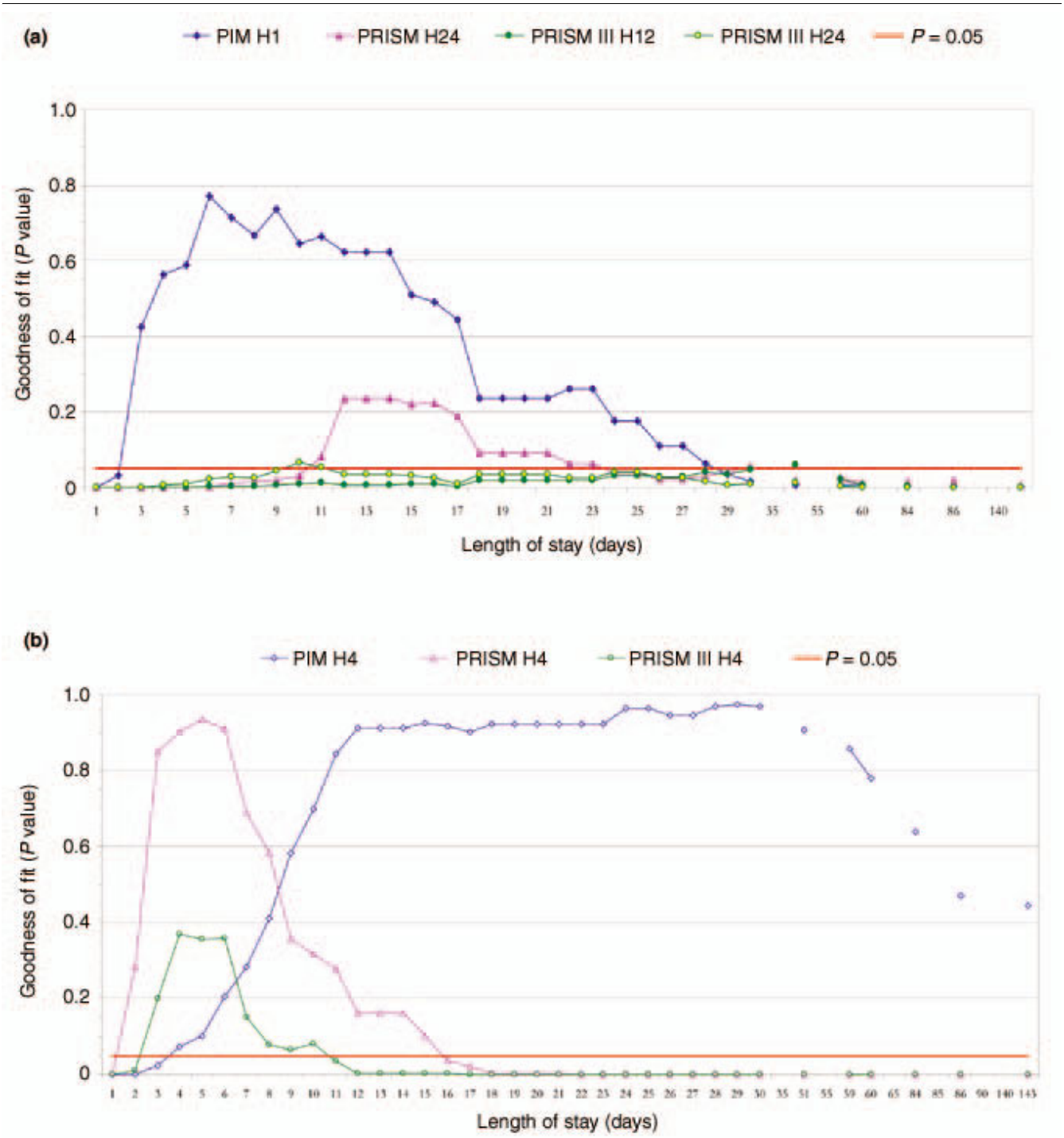
0.14 for the PRISM III 12-hour assessment, 0.04 (indicating no statistical calibration) for the PRISM III 24-hour assessment, and 0.07 for the PIM score. Pearson and coworkers [18] tested the PIM score in a PICU population of 7253 children; the χ^2 goodness-of-fit test value calculated from their data was 37.4 ($P < 0.0001$, df = 10), which suggests no statistical calibration, as indicated by others [19-21]. Tibby and coworkers [5] compared the pre-ICU PRISM, PIM and PRISM scores in 928 patients. They concluded that all scoring systems exhibited suboptimal calibration ($P = 0.08$ for the PRISM, and $P < 0.0001$ for the pre-ICU PRISM and PIM). Slater and coworkers [16] observed 20 PICU deaths in their study, including 598 children from one unit (21 with inclusion criteria for the PRISM III, which considers patients who die within 24 hours of PICU discharge), whereas expected deaths were 21.3, 32.3 and 23.4 for the PIM, PRISM and PRISM III scores, respectively. Although goodness-of-fit test values were not calculated in their study [16], calibration of the PRISM score could be expected to be poor.

The previously reported miscalibration of the PRISM score [22,23] led Tilford and coworkers [24] to use a different set of coefficient estimates. When interpreting the calibration of the PRISM III score, the version selected must be considered. In the present study the PRISM III score was calculated using the 1996 version and not the 1999 one, which includes other variables that are not described in the first PRISM III report and, to our knowledge, have not been reported elsewhere [2].

In our study, as in that by Gemke and van Vught [15], the expected mortality underestimated the observed mortality in the group at low risk for mortality and overestimated it in the group at very high risk for mortality (>30%). Such discrepancies have been reported with both paediatric [23] and adult [25] generic scoring systems.

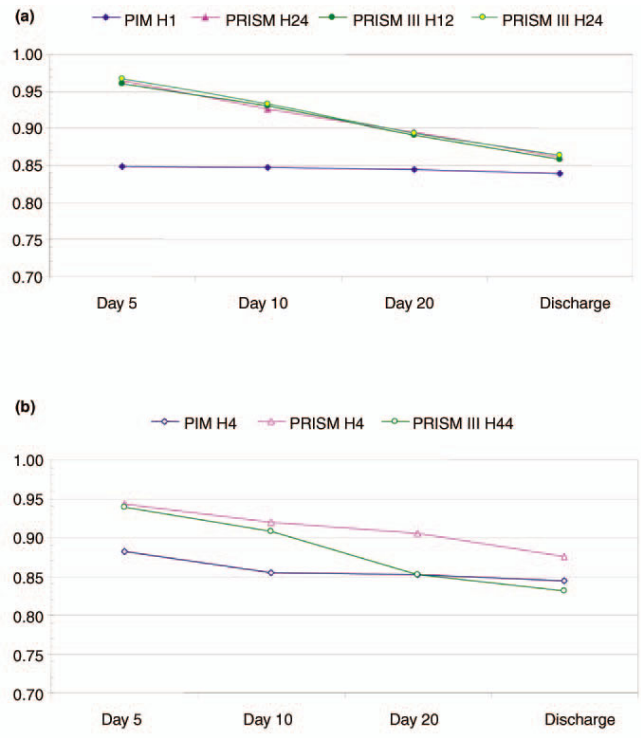
The length of stay was studied by Bertolini and coworkers [23] because the PRISM score could not correctly predict outcome. Those authors found a good calibration for patients with a length of stay of 4 days or less and a poor calibration in those patients who stayed for longer than 4 days. The present study showed that, for the time periods for which the scores were developed, the PIM score provided the earliest (from day 3) and longest (to day 28) calibration. For a different time point

Figure 1



Effect of the length of stay on calibration of the Paediatric Index of Mortality (PIM) [4], Paediatric RISk of Mortality (PRISM) and PRISM III scores [1,2] for **(a)** the time points for which the scores were developed (1 hour [H1] for PIM, 24 hours [H24] for PRISM, and 12 hours [H12] and H24 for PRISM III) and **(b)** a different time period, namely 4 hours (H4).

Figure 2



Effect of the length of stay on discrimination of the Paediatric Index of Mortality (PIM) [4], Paediatric RISk of Mortality (PRISM) and PRISM III scores [1,2] for **(a)** the time points for which the scores were developed (1 hour [H1] for PIM, 24 hours [H24] for PRISM, and 12 hours [H12] and H24 for PRISM III) and **(b)** a different time period, namely 4 hours (H4).

(i.e. 4 hours), the three scores were calibrated after a few days: day 2 for the PRISM 4-hour assessment, day 3 for the PRISM III 4-hour assessment, and day 4 for PIM 4-hour assessment; only the PIM 4-hour assessment was calibrated until discharge.

Moreover, patient mortality is affected by demographical, physiological and diagnostic data, but it also depends on many other factors such as comorbidities, which did not appear to be accounted for sufficiently in our population. In the recently reported PIM2 [3], the numbers of diagnostic criteria (high risk and low risk diagnosis) and comorbidities have been increased. Discrepancies between discrimination and calibration have previously been discussed. In fact, PRISM score, Acute Physiology and Chronic Health Evaluation (APACHE) score, Mortality Probability Model (MPM) score and Simplified Acute Physiology Score (SAPS) were reported in several studies to exhibit good discrimination but poor calibration [23,25-29]. Unsatisfactory calibration of scores can be attributed to various factors, including poor performance of the medical system (if observed mortality is greater than predicted mortality) [23,25], differences in case mix [27] and mortality rate [30], as well as failure of the score equation to model the actual situation accurately [25].

The above mentioned paediatric studies did not give any information on the childrens' characteristics (case mix), which potentially could explain discrepancies between discrimination and calibration [2,15,23]. Indeed, the two studies using the additional variables of the PRISM III score [2,15] did not provide a clear description of their population. Important differences in case mix data are represented by mortality rates, which were different between PICUs (e.g. 4.8% for Pollack and coworkers [2], 6.6% for Gemke and van Vught [15] and 10.0% in the present study). The further the hospital mortality rate diverged from the original rate, the worse the performance of the model [17]. Goodness-of-fit tests are more sensitive than AUCs [17], and it has been suggested that, in the presence of good discrimination, bad calibration due to the source is correctable by using customization [31,32]. However, Diamond [33] demonstrated that perfect calibration and perfect discrimination cannot coexist; a perfectly calibrated model is not perfectly discriminatory because it has an AUC of only 0.83 rather than 1. Customization of a score is justified when the database on which it was developed is old and when the score is used in a specific population [24]. However, customization by a unit could lead to inability to evaluate (or compare) performance between units.

Is a score with poor calibration useful? If scores are used to assess quality of care, as estimated by SMR, then calibration, rather than discrimination, is the best measure of performance. It is also recognized that there are no formal means of directly comparing the χ^2 values derived from the goodness-of-fit test [30]. Our data and those reported by Livingston and coworkers [30] showed large differences in χ^2 goodness-of-fit test values between several scores. Thus, one can consider that a way to describe calibration of a score is to detail the χ^2 goodness-of-fit test values for different classes of mortality probability, which reflects exact prediction across the full range of severity (Tables 3 and 5) [18,20].

Stratification for inclusion of children in clinical trials remains an important problem in PICUs [6]. Scoring systems are used to compare or control for severity of illness in clinical trials and have been integrated into guidelines [6]. The question is, what kind of scoring system do we need if we are to include children in clinical trials? We probably need a score that represents well the patient's condition early after admission to the PICU. With this aim in mind, the PIM score appears superior to the PRISM and PRISM III scores. PIM score takes into account the condition of the patient directly on arrival in the PICU (i.e. when the patient's condition is least affected by therapeutic intervention). PRISM score require an observation period of 24 hours, which represents a limitation of its use as an inclusion criterion in clinical trials. To date, no consensus has been reached as to which approach represents the 'gold standard' [7]. In order

to minimize inclusion delay, Pollack and coworkers [2] proposed estimation of the probability of death using the PRISM III calculated 12 hours after admission. However, this delay is too long for serious diseases (e.g. meningococcal septic shock). In the present study the performance of PIM at 4 hours was better than at 1 hour. Thus, a 4-hour observation period seems to be a good compromise, allowing evaluation of the patient's clinical condition and permitting stabilization, without delaying inclusion in a therapeutic trial. We arbitrarily chose a period of 4 hours after PICU admission. Calculation of the scores at 3 or 5 hours would probably have yielded similar results.

To our knowledge, no study has compared the performance of generic paediatric mortality scores calculated within a few hours of admission to the PICU. Castellanos-Ortega and coworkers [9] used a similar approach in a specific population of children with meningococcal septic shock by calculating one generic (PIM) and two specific scores 2 hours after PICU admission; the PIM 2-hour score was as discriminant (AUC 0.82) as their new score (AUC 0.92; $P = 0.10$) but exhibited poor calibration.

Conclusion

The present study indicates that, among generic scores calculated at 4 hours after admission and with good discriminatory capacity (i.e. AUC > 0.80), only the PIM 4-hour score was well calibrated. The updated PIM2, which takes into account new primary reasons for ICU admission and comorbidities, must be validated for the time point for which it was developed and at a different time point. Further studies are required before the PIM (or PIM2) 4-hour score can be used as an inclusion criterion for clinical trials.

Key messages

Among generic paediatric mortality scores calculated at 4 hours after admission in 802 consecutive children, only the PIM score was both discriminant and Calibrated

Competing interests

None declared.

Acknowledgements

The authors' contributions were as follows: study conception and design: Francis Leclerc and Stéphane Leteurtre; acquisition of data: Stéphane Leteurtre, Jessica Wirth, Odile Noizet, Eric Magnenant, Ahmed Sadik, Catherine Fourier and Robin Cremer; Analysis and interpretation of data: Stéphane Leteurtre, Francis Leclerc and Jessica Wirth; Draft of the article: Stéphane Leteurtre, Francis Leclerc and Eric Magnenant; critical revision of the manuscript for important intellectual content: all investigators read and commented regarding important intellectual content; statistical expertise: Stéphane Leteurtre and Jessica Wirth; Admistrative, technical, or materiel support: Stéphane Leteurtre, Ahmed Sadik, Odile Noizet, Catherine Fourier and Robin Cremer; supervision: Francis Leclerc.

References

1. Pollack MM, Ruttimann UE, Getson PR: **Pediatric risk of mortality (PRISM) score.** *Crit Care Med* 1988, **16:**1110-1116.
2. Pollack MM, Patel KM, Ruttimann UE: **PRISM III: an updated Pediatric Risk of Mortality score.** *Crit Care Med* 1996, **24:**743-752.
3. Slater A, Shann F, Pearson G: **PIM2: a revised version of the Paediatric Index of Mortality.** *Intensive Care Med* 2003, **29:**278-285.
4. Shann F, Pearson G, Slater A, Wilkinson K: **Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care.** *Intensive Care Med* 1997, **23:**201-207.
5. Tibby SM, Taylor D, Festa M, Hanna S, Hatherill M, Jones G, Habibi P, Durward A, Murdoch IA: **A comparison of three scoring systems for mortality risk among retrieved intensive care patients.** *Arch Dis Child* 2002, **87:**421-425.
6. Marcin JP, Pollack MM: **Review of the methodologies and applications of scoring systems in neonatal and pediatric intensive care.** *Pediatr Crit Care Med* 2000, **1:**20-27.
7. Randolph AG: **Paediatric index of mortality (PIM): do we need another paediatric mortality prediction score?** *Intensive Care Med* 1997, **23:**141-142.
8. Rello J, Rue M, Jubert P, Muses G, Sonora R, Valles J, Niederman MS: **Survival in patients with nosocomial pneumonia: impact of the severity of illness and the etiologic agent.** *Crit Care Med* 1997, **25:**1862-1867.
9. Castellanos-Ortega A, Delgado-Rodriguez M, Llorca J, Sanchez Buron P, Mencia Bartolome S, Soult Rubio A, Milano Manso G, Dominguez Sampedro P, Blanco Montero R, Rodriguez Nunez A, Zambrano Perez E, Rey Galan C, Lopez Negueruela N, Reig Saenz R: **A new prognostic scoring system for meningococcal septic shock in children. Comparison with three other scoring systems.** *Intensive Care Med* 2002, **28:**341-351.
10. Lemeshow S, Hosmer DW: **A review of goodness of fit statistics for use in the development of logistic regression models.** *Am J Epidemiol* 1982, **115:**92-106.
11. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143:**29-36.
12. Murphy-Filkins R, Teres D, Lemeshow S, Hosmer DW: **Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit.** *Crit Care Med* 1996, **24:**1968-1973.
13. Breslow NE, Day NE: **The design and analysis of cohort studies.** In *Statistical Methods in Cancer Research. International Agency for Research on Cancer, Scientific Publication no. 82* 1987:61-64.
14. Randolph AG, Guyatt GH, Calvin JE, Doig G, Richardson WS: **Understanding articles describing clinical prediction tools. Evidence Based Medicine in Critical Care Group.** *Crit Care Med* 1998, **26:**1603-1612.
15. Gemke RJ, van Vught J: **Scoring systems in pediatric intensive care: PRISM III versus PIM.** *Intensive Care Med* 2002, **28:**204-207.
16. Slater AJ, Dempster MC, Fyfe JA, Jones SR: **Comparison of mortality prediction using PIM, PRISM II and PRISM III-24 in an Australin pediatric intensive care unit.** *Intensive Care Med* 1998, **24:**S186.
17. Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D: **Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study.** *Crit Care Med* 1996, **24:**57-63.
18. Pearson GA, Stickley J, Shann F: **Calibration of the paediatric index of mortality in UK paediatric intensive care units.** *Arch Dis Child* 2001, **84:**125-128.
19. Parry G, Jones S, Simic-Lawson M: **Calibration of the paediatric index of mortality in UK paediatric intensive care units.** *Arch Dis Child* 2002, **86:**65-66.
20. Shann F: **Are we doing a good job: PRISM, PIM and all that.** *Intensive Care Med* 2002, **28:**105-107.

21. Tibby SM, Murdoch IA: **Calibration of the paediatric index of mortality score for UK paediatric intensive care.** *Arch Dis Child* 2002, **86:**65-66.
22. Goddard JM: **Pediatric risk of mortality scoring overestimates severity of illness in infants.** *Crit Care Med* 1992, **20:**1662-1665.
23. Bertolini G, Ripamonti D, Cattaneo A, Apolone G: **Pediatric risk of mortality: an assessment of its performance in a sample of 26 Italian intensive care units.** *Crit Care Med* 1998, **26:**1427-1432.
24. Tilford JM, Roberson PK, Lensing S, Fiser DH: **Differences in pediatric ICU mortality risk over time.** *Crit Care Med* 1998, **26:**1737-1743.
25. Pappachan JV, Millar B, Bennett ED, Smith GB: **Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system.** *Chest* 1999, **115:**802-810.
26. Bastos PG, Sun X, Wagner DP, Knaus WA, Zimmerman JE: **Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study.** *Intensive Care Med* 1996, **22:**564-570.
27. Markgraf R, Deutschinoff G, Pientka L, Scholten T: **Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit.** *Crit Care Med* 2000, **28:**26-33.
28. Nouira S, Belghith M, Elatrous S, Jaafoura M, Ellouzi M, Boujdaria R, Gahbiche M, Bouchoucha S, Abroug F: **Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units.** *Crit Care Med* 1998, **26:**852-859.
29. Wood KE, Coursin DB, Grounds RM: **Critical care outcomes in the United Kingdom: sobering wake-up call or stability of the lamppost?** *Chest* 1999, **115:**614-616.
30. Livingston BM, MacKirdy FN, Howie JC, Jones R, Norrie JD: **Assessment of the performance of five intensive care scoring models within a large Scottish database.** *Crit Care Med* 2000, **28:**1820-1827.
31. Metnitz PG, Lang T, Vesely H, Valentin A, Le Gall JR: **Ratios of observed to expected mortality are affected by differences in case mix and quality of care.** *Intensive Care Med* 2000, **26:**1466-1472.
32. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA: **Regression modelling strategies for improved prognostic prediction.** *Stat Med* 1984, **3:**143-152.
33. Diamond GA: **What price perfection? Calibration and discrimination of clinical prediction models.** *J Clin Epidemiol* 1992, **45:**85-89.