Citation	Jacobs A, Flechet M, (2019), Performance of Pediatric Mortality Prediction Scores for PICU Mortality and 90-Day Mortality Pediatr Crit Care Med. 2019 Feb;20(2):113-119
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	http://dx.doi.org/10.1097/PCC.0000000000001764
Journal homepage	Pediatric Critical Care Medicine
Author contact	greet.vandenberghe@kuleuven.be + 32 (0)16 34 40 21

(article begins on next page)

Performance of pediatric mortality prediction scores for pediatric intensive care unit mortality and 90-day mortality

An Jacobs MD^{1*}, Marine Flechet MEng^{1*}, Ilse Vanhorebeek PhD¹, Sören Verstraete MD¹, Catherine Ingels MD¹, Michael P Casaer MD¹, Gerardo Soto-Campos PhD², Sascha C Verbruggen MD³, Koen F Joosten MD³, Fabian Güiza PhD¹, Greet Van den Berghe MD^{1C}

¹Clinical Division and Laboratory of Intensive Care Medicine, Department of Cellular and Molecular Medicine, KU Leuven University Hospital, Leuven, Belgium.

²Virtual Pediatric Systems, LCC, Los Angeles, California, United States.

³Department of Pediatrics and Pediatric Surgery, Erasmus Medical Centre, Sophia Children's Hospital, Rotterdam, The Netherlands.

* Equal contribution.

^C Corresponding author.

Author's contact information

- An Jacobs – Herestraat 49 bus 503, 3000 Leuven, Belgium. +32 16 34 09 67. an.jacobs@kuleuven.be
- Marine Flechet – Herestraat 49 bus 503, 3000 Leuven, Belgium. +32 16 33 02 68. marine.flechet@kuleuven.be
- Ilse Vanhorebeek – Herestraat 49 bus 503, 3000 Leuven, Belgium. +32 16 33 02 68. ilse.vanhorebeek@kuleuven.be
- Sören Verstraete – Herestraat 49 bus 503, 3000 Leuven, Belgium. +32 16 33 02 68. soren.verstaete@kuleuven.be
- Catherine Ingels – Herestraat 49 bus 7003, 3000 Leuven, Belgium. +32 16 34 40 21. catherine.ingels@uzleuven.be
- Michael P Casaer – Herestraat 49 bus 7003, 3000 Leuven, Belgium. +32 16 34 40 21. michael.casaer@uzleuven.be
- Gerardo Soto-Campos - 4470 W. Sunset Blvd, 440, Los Angeles, California, USA. +1 888 999 4850. gsotocampos@myvps.org
- Sascha C Verbruggen – Dr Molewaterplein 60, 3015 GJ Rotterdam, The Netherlands. +31 10 70 40 704. s.verbruggen@erasmusmc.nl
- Koen F Joosten – Dr Molewaterplein 60, 3015 GJ Rotterdam, The Netherlands. +31 10 70 40 704. k.joosten@erasmusmc.nl
- Fabian Güiza – Herestraat 49 bus 503, 3000 Leuven, Belgium. +32 16 33 02 68. fabian.guiza@kuleuven.be
- Greet Van den Berghe – Herestraat 49 bus 7003, 3000 Leuven, Belgium. +32 16 34 40 21. greet.vandenbergh@uzleuven.be (corresponding author)

No reprints will be ordered.

Author Disclosure Statement

The authors report no conflicts of interest in this work. Virtual Pediatric Systems has calculated PRISM III probabilities of death, but has not been involved in data analysis, nor manuscript preparation.

Funding

M. Flechet and S. Verstraete received funding from the Research Foundation, Flanders (FWO) as a PhD fellow. C. Ingels received funding from the Clinical Research Foundation from the University Hospitals Leuven, Belgium. M. Casaer received support from a postdoctoral research grant by FWO. S. Verbruggen received funding from Fonds NutsOhra, Sophia Research Foundation (SSWO), Stichting Agis Zorginnovatie and the Erasmus Trustfonds. I. Vanhorebeek and G. Van den Berghe received funding from the Methusalem program of the Flemish Government (Belgium, METH14/06). G. Van den Berghe further received funding from the Flemish Agency for Innovation through Science and Technology (IWT-TBM110685) and the European Research Council (Advanced Grant AdvG-2012–321670).

Copyright form disclosure: Dr. Verstraete and M. Flechet received funding from Research Foundation Flanders (FWO). Dr. Ingels' institution received funding from clinical research foundation of UZLeuven and a doctoral fellowship of FWO Belgium. Dr. Casaer received support from a postdoctoral research grant by FWO. Dr Verbruggen received funding from Fonds NutsOhra, Sophia Research Foundation (SSWO), Stichting Agis Zorginnovatie and the Erasmus Trustfonds. I. Vanhorebeek received funding from the Methusalem program of the Flemish Government (Belgium, METH14/06). Dr Van den Berghe received funding from the Methusalem program of the Flemish Government (Belgium, METH14/06), Flemish Agency for Innovation through Science and Technology (IWT-TBM110685) and the European Research Council (Advanced Grant AdvG-2012-321670). The remaining authors have disclosed that they do not have any potential conflicts of interest

Key words:

Mortality prediction scores, pediatric critical care, PICU, clinical trials, 90-day mortality, outcome.

Abstract

Objective: The use of mortality prediction scores in clinical trials in the pediatric-intensive-care-unit (PICU) is essential for comparing patient groups. Because of the decline in PICU mortality over the last decades, leading to a shift towards later deaths, recent trials use 90-day mortality as primary outcome for estimating mortality and survival more accurately. This study assessed and compared the performance of two frequently used PICU mortality prediction scores for prediction of PICU and 90-day mortality.

Design: This secondary analysis of the randomized controlled PEPaNIC trial compared the discrimination (area under the receiver-operating characteristic curve, AUC) and calibration of the pediatric index of mortality (PIM) 3 and the pediatric risk of mortality (PRISM) III scores for prediction of PICU- and 90-day mortality.

Setting: Three participating PICUs within academic hospitals in Belgium, the Netherlands and Canada.

Patients: 1428 critically ill patients aged 0-17 years.

Interventions: None.

Main results: Although PIM3 only includes information available at the time of PICU admission, thus before any intervention in the PICU, it showed good discrimination (AUC 0.894, 95%CI 0.892-0.896), and good calibration (no deviation from the diagonal, $p=0.58$) for PICU mortality. PRISM III, which involves the worst values for the evaluated variables during the first 24 hours of PICU stay, was statistically more discriminant (AUC 0.920, 95%CI 0.918-0.921, $p=0.04$) but poor in calibration (significant deviation from the diagonal, $p=0.04$). PIM3 and PRISM III discriminated equally well between 90-day mortality and survival (AUC 0.867, 95%CI 0.866-0.869 and AUC 0.882, 95%CI 0.880-0.884, respectively, $p=0.77$), but PRISM III was not well calibrated ($p=0.04$), unlike PIM3 ($p=0.34$).

Conclusion: PIM3 performed better in calibration for predicting PICU and 90-day mortality than PRISM III and is not influenced by intervention or PICU quality of care. Therefore, PIM3 seems a better choice for use in clinical trials with 90-day mortality as primary outcome.

Article Tweet

Mortality prediction scores used in clinical trials in the PICU with 90-day mortality as primary outcome: PIM3 seems to be the better choice

Introduction

For over 3 decades, scores have been developed to predict risk of death for critically ill infants and children while being treated in the pediatric intensive care unit (PICU) (1). The purpose of these mortality prediction scores is threefold. First and foremost, they serve to compare the standard of care between and within units over time. Second, they enable comparison of different organization systems of intensive care. Finally, they allow comparison between groups of patients in research trials (2). Mortality prediction scores are not intended for, and are therefore not appropriate for, individual patient decision-making (3,4). Since patients in clinical trials need to be comparable at baseline regarding their mortality risk, an adequate mortality prediction score should show good concordance between observed and predicted mortality. In addition, scoring should be performed as soon as possible after PICU admission, to exclude influence of any intervention during admission.

Currently, two mortality prediction scores (with various versions) are being used in the PICU. The Pediatric Index of Mortality (PIM) score was developed from data collected between 1994 and 1996 in PICUs from Australia and the United Kingdom (5). It has been revised into the PIM2 score in 2003 to adjust for improvements in PICU outcomes (2), and into the PIM3 score in 2013 to improve calibration (6). The Pediatric Risk of Mortality (PRISM) score was developed by deriving data from PICUs in the United States between 1980 and 1985 (7). A revision to refine the original score in 1996 to adapt for improved care and patient population changes resulted in the PRISM III (8).

Two major caveats need to be considered when using mortality prediction scores in the PICU. First, during the last decades an important decline in PICU mortality has been observed worldwide (9-11), due to new insights in diagnostic and therapeutic measures in the field of pediatric critical care medicine (12), centralization of care (13), and specialized staff training and education (14,15). In addition, the number of critically ill children who suffer from an underlying chronic complex illness is increasing (11,16). Therefore, the scores have to be updated regularly to maintain good performance (17). Second, PICU mortality is a location-dependent mortality endpoint, predicted by the existing

mortality prediction scoring systems and easy to collect, but could be influenced by discharge policies, number of PICU beds available, hospital characteristics and geographical differences (18). A duration-dependent mortality endpoint, such as 90-day mortality, is independent of those logistics, although it might be less practical to collect these data (18,19). Recent studies in adults after oesophagectomy have shown that longer follow-up periods up until 90 days may be necessary to assess mortality, since the intensive care in the acute phase improved and consequently led to a shift towards later deaths (20,21). In addition, in recently published trials in ICU, there is also a shift from ICU mortality to 90-day mortality as a primary endpoint (22-25), albeit without comparison of the models. So far, there is no consensus yet on which is the best option as a primary mortality endpoint for critically ill children.

Nevertheless, there is a need for the development of scores specifically designed for the prediction of 90-day mortality in critically ill children. The aim of this study was to assess and compare the performance of the existing and widely used scores for prediction of PICU mortality as their original endpoint, and to evaluate whether the use of these scores could be extended to prediction of 90-day mortality. Therefore, we assessed and compared discrimination and calibration of the PIM3 and PRISM III scores.

Methods

Study population

This is a secondary analysis of the multicenter randomized controlled PEPaNIC trial (26), which assessed the impact of early nutritional management in 1440 critically ill children. Children aged 0 to 17 years were included if a stay of 24 hours or more in the PICU was expected. Patients who were assessed for eligibility but who were not included in the PEPaNIC trial were mostly not ill enough to necessitate nutritional support, had a STRONGkids score (Screening Tool for Risk of Nutritional Status and Growth, indicating the risk of malnutrition) of less than 2, or were not included because of a

readmission to the PICU. The Institutional Ethical Review Boards of the participating centers in Leuven, Belgium (ML8052), Rotterdam, The Netherlands (NL38772.000.12) and Edmonton, Canada (Pro00038098) approved the study, which was performed in accordance with the 1964 Declaration of Helsinki and its amendments. Written informed consent for data analyses was obtained from the parents or legal guardians.

Mortality prediction scores

The most recently developed and externally validated versions of the PIM and PRISM scores, PIM3 and PRISM III, were calculated (6,8). Scoring for PIM3 was performed retrospectively but blinded for outcome, while PRISM III was scored prospectively during PICU admission. Probabilities of death based on the PIM3 score were calculated with the published regression formula (6). As the equivalent regression formula for PRISM III is not part of the public domain, probability of death for PRISM III was calculated with the original equation by Virtual Pediatric Systems (Los Angeles, CA, USA; <http://www.myvps.org>). Differences between PIM3 and PRISM III are listed in table 1.

Endpoints

Predictions of PICU mortality, the original endpoint for which the scores were developed, and mortality at 90 days after PICU admission were evaluated.

Statistical analysis

Logistic regression was used to investigate the independent association of each score with the endpoints. One thousand bootstrap samples were used to create confidence intervals and evaluate the stability of each model, and to compare areas under the receiver-operating characteristics (ROC) curves between scores (27).

Model performance was assessed through discrimination and calibration analysis (28). Discrimination is the ability of the score to distinguish between mortality and survival, and was reported using receiver-operating characteristics (ROC) curves and area under the ROC curves (AUC) (29). An AUC

above 0.7 is considered to reflect adequate discrimination (30). Calibration refers to the agreement between the observed mortality and the predicted mortality by the score (31). For example, if the observed frequency of mortality is 30 out of 100 patients, a well-calibrated score should predict a 30% risk of mortality for these 100 patients. Calibration was visualized using the GiViTI calibration belt (32) together with patient distributions stratified by endpoint (29), and quantified using the calibration test (31). Adequate calibration refers to a calibration curve close to the diagonal. A statistically significant difference from the diagonal is reported by a calibration test with a P-value < 0.05.

Statistical analyses were performed with JMP® 13.0.0 (SAS Institute, Inc, Cary, NC), Python version 2.7.8 (Python Software Foundation, <http://www.python.org>), and Scipy version 0.15 (SciPy.org). Statistical significance was set at $P < 0.05$.

Results

Of the 1440 patients included in the PEPaNIC trial, 12 were excluded because of missing blood gas variables to calculate the PRISM III score. The baseline characteristics of the remaining 1428 patients are presented in table 2.

PICU Mortality

Out of the 1428 patients, 68 died during PICU admission (4.8%). PIM3 and PRISM III both discriminated well between PICU mortality and survival (Figure 1A mean AUC 0.894 (95% CI 0.892-0.896) and 0.920 (95% CI 0.918-0.921), respectively), but PRISM III discriminated statistically better than PIM3 ($p=0.04$). However, PIM3 showed good calibration (Figure 2, $p=0.58$), in contrast to PRISM III, which deviated significantly from the diagonal ($p=0.04$). PIM3 calibrated well for all risk ranges, whereas PRISM III was well calibrated for middle to high probabilities, but underestimated lower risks of mortality.

90-day mortality

After 90 days, 87 (6.1%) patients had died. PIM3 and PRISM III both discriminated well and comparably between 90-day mortality and survival (mean AUC 0.867 (95% CI 0.866-0.869) and 0.882 (95% CI 0.880-0.884), respectively, $p=0.77$) (figure 1B). However, discrimination of both PIM3 and PRISM III was significantly lower for 90-day mortality than for PICU mortality ($p=0.01$ and $p=0.002$ respectively). Concerning calibration, the calibration belt for PIM3 did not deviate from the diagonal ($p=0.34$, figure 3), whereas PRISM III again underestimated the risk of 90-day mortality in the lower risk ranges ($p=0.04$).

Discussion

We compared the performance between two pediatric mortality prediction scores for both PICU mortality and for 90-day mortality in a multicenter cohort of a subgroup of critically ill children with different underlying medical conditions that required admission to a PICU. PRISM III was significantly better in discriminating between PICU mortality and survival as compared with PIM3. There was no difference between the scores in discrimination between 90-day mortality and survival, an endpoint for which the scores had not been designed originally. Finally, for calibration for both endpoints, PIM3 performed well, whereas PRISM III did not.

For discrimination between PICU mortality and survival, the endpoint for which the scores were designed, both scores perform very well, confirming the results of their development studies. The AUC of PIM3 in our study is similar to the AUC of its development cohort (AUC 0.88) (6).

Discrimination of PRISM III also reflected the initial results of the development study (AUC of 0.959) (8), and appeared to be significantly better in discriminating PICU mortality versus survival as compared to PIM3. However, despite its much shorter period of data collection, PIM3 still discriminated very well. Importantly, in terms of calibration, PIM3 outperformed PRISM III. Calibration of PRISM III was suboptimal for low mortality risk categories, where it substantially

underestimated death. This contrasts with its development study, where it was well calibrated (8). This might be explained by the fact that data of its development study were collected in PICUs in the United States only, which might cause demographic bias when applying the score elsewhere, and by the fact that our patient cohort is a subgroup of the admitted PICU patients. In addition to the relatively small number of patients, our patient group might not be representative as compared with the ones from the development and validation studies. Finally, the development study uses the Hosmer-Lemeshow goodness-of-fit analysis for evaluating calibration (8). Calibration in our study was not evaluated by this analysis, since it has poor power to detect miscalibration of systematic differences between outcomes in the new data and the model development data, or to detect overfitting (33).

Both PIM3 and PRISM III performed very well in predicting 90-day mortality, which makes them suitable for use in clinical trials with 90-day mortality as an endpoint. There was no significant difference between both scores in discriminating survival and mortality at 90 days. However, PRISM III underestimated the mortality risk for low probabilities. PIM3's better calibration for both PICU and 90-day mortality might be due to the recent revision of PIM2 into PIM3 to improve its calibration, and because of the fact that PRISM III was developed much earlier in time.

As listed in table 1, the scores differ in the parameters required to calculate the score, the starting point and duration of the observation period, and the equation to calculate the probability of death from the score (6,8). PRISM III variables are collected during the first 24 hours of PICU admission, whereas PIM3 evaluates the patient's condition directly at PICU admission. Consequently, PRISM III variables can theoretically be influenced by diagnostic and therapeutic measures during the 24 hours of PICU admission, and can therefore obscure poor quality of care (34). In addition, a relatively large proportion of the non-survivors die during the first 24 to 48 hours of admission. Hence, during this time period, PRISM III is diagnosing death rather than predicting it (35). This causes methodological concerns for PRISM III. In addition, PIM3 might be easier to use in daily practice, since collection of

the variables takes place at one time point, and the equation for calculating the probability of death based on the PIM3 score is part of the public domain.

Recently, Pollack et al revised PRISM III with development of PRISM IV (36). Their goal was to adapt the score to new practice patterns, to minimize bias and to reduce potential sources of error. For the calculation of PRISM IV, the window of data collection is shortened to 2 hours before PICU admission to 4 hours after admission for laboratory data, and to the first 4 hours of PICU management for other variables. Additionally, the equation to calculate probability of death based on the PRISM IV score is part of the public domain, which is a significant improvement in accessibility and transparency as compared to PRISM III. The most innovative difference between the new score with respect to PRISM III and PIM3, is the change to hospital mortality as endpoint. Hospital mortality might be a more relevant prediction target since it has a higher rate than PICU mortality, but should be further validated in future studies. However, the duration from admission up to hospital discharge or mortality is very heterogeneous in patient cohorts, and the timing of hospital discharge depends on local practices. Instead of hospital mortality, a fixed time point such as 90-day mortality, which is independent of discharge status and thus eliminates the competing event of discharge from mortality analysis, is deemed a more appropriate endpoint for use in clinical trials (22-25). Unfortunately, for our study, the appropriate data to calculate PRISM IV were not collected prospectively since the revision was published after PEPaNIC inclusion (26,36), and could not be traced retrospectively.

Our study has some limitations that need to be addressed. First, calculation of the PIM3 score was done retrospectively, which might be more difficult to detect errors in the variables. However, all laboratory and physiological variables were collected prospectively in the context of the original clinical trial. Only variables concerning diagnostic risk categories were determined retrospectively.

Second, as compared to the development and validation studies of the scores, our cohort of 1428 critically ill patients (with 68 and 87 deaths during PICU stay and at 90 days respectively) is small and represents a subgroup of the critically ill children admitted to the PICUs that participated in the PEPaNIC trial. Larger external validation trials are needed to confirm our results. However, we used a robust statistical methodology with 1000 bootstraps to assess the generalizability and robustness of our results, which are reflected by tight confidence intervals. Strengths of this study are the multicenter cohort of critically ill children with heterogeneous medical conditions underlying the necessity for PICU admission. Additionally, we used advanced statistical analysis including bootstrapping to improve robustness of the results, and calibration belts with visualization of distribution of the patients to overcome the risk of overfitting and poor detection of miscalibration. Finally, to the best of our knowledge, this is the first study to compare mortality prediction scores for 90-day mortality prediction, which has become the most relevant endpoint for clinical trials.

Conclusion

PIM3 and PRISM III, developed to predict PICU mortality, both nevertheless performed well in terms of discrimination between 90-day mortality and survival. This is an important finding for the use of the scores in clinical trials with 90-day mortality as primary endpoint. PIM3 appeared to be a better choice for use in clinical trials with 90-day mortality as primary endpoint as compared to PRISM III, because of better calibration, a superior methodology by scoring at admission and therefore before intervention, and because its equation is part of the public domain. However, our findings should be confirmed in larger external validation studies, and ideally, a new score specifically designed to predict 90-day mortality should be developed.

Acknowledgments

We thank the research nurses and nurses and doctors of the PICU, who collected, entered and cleaned the data. We also wish to acknowledge the efforts of Pieter J Wouters for the database support.

References

1. Pollack MM, Yeh TS, Ruttiman UE, et al. Evaluation of pediatric intensive care. *Crit Care Med.* 1984;12(4):376-383.
2. Slater A, Shann F, Pearson G, et al. PIM2: a revised version of the Paediatric Index of Mortality. *Intensive Care Med.* 2003;29(2):278-285.
3. Vincent JL, Opal SM, Marshall JC. Ten reasons why we should NOT use severity scores as entry criteria for clinical trials or in our treatment decisions. *Crit Care Med.* 2010;38(1):283-287.
4. Wong HR. Estimating Mortality Risk of Pediatric Critical Illness: A Worthy Obsession. *Pediatr Crit Care Med.* 2016;17(9):887-888.
5. Shann F, Pearson G, Slater A, et al. Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care. *Intensive Care Med.* 1997;23(2):201-207.
6. Straney L, Clements A, Parslow RC, et al. Paediatric index of mortality 3: an updated model for predicting mortality in pediatric intensive care*. *Pediatr Crit Care Med.* 2013;14(7):673-681.
7. Pollack MM, Ruttimann UE, Getson PR. Pediatric risk of mortality (PRISM) score. *Crit Care Med.* 1988;16(11):1110-1116.
8. Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated Pediatric Risk of Mortality score. *Crit Care Med.* 1996;24(5):743-752.
9. Namachivayam P, Shann F, Shekerdemian L, et al. Three decades of pediatric intensive care: Who was admitted, what happened in intensive care, and what happened afterward. *Pediatr Crit Care Med.* 2010;11(5):549-555.
10. Burns JP, Sellers DE, Meyer EC, et al. Epidemiology of death in the PICU at five U.S. teaching hospitals*. *Crit Care Med.* 2014;42(9):2101-2108.

11. Rennick JE, Childerhose JE. Redefining success in the PICU: new patient populations shift targets of care. *Pediatrics*. 2015;135(2):e289-291.
12. Carcillo JA. What's new in pediatric intensive care. *Crit Care Med*. 2006;34(9 Suppl):S183-190.
13. Pearson G, Shann F, Barry P, et al. Should paediatric intensive care be centralised? Trent versus Victoria. *Lancet*. 1997;349(9060):1213-1217.
14. Gupta P, Tang X, Rettiganti M, et al. Association of house staff training with mortality in children with critical illness. *Acta Paediatr*. 2016;105(2):e60-66.
15. Pollack MM, Patel KM, Ruttimann E. Pediatric critical care training programs have a positive effect on pediatric intensive care mortality. *Crit Care Med*. 1997;25(10):1637-1642.
16. Edwards JD, Houtrow AJ, Vasilevskis EE, et al. Chronic conditions among children admitted to U.S. pediatric intensive care units: their prevalence and impact on risk for mortality and prolonged length of stay*. *Crit Care Med*. 2012;40(7):2196-2203.
17. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ*. 2015;351:h3868.
18. Groeneveld JA. Mortality as an endpoint in studies in critically ill patients: a reappraisal of definitions. *Minerva Anesthesiol*. 2016;82(7):785-790.
19. Suffredini AF. A Role for Hydrocortisone Therapy in Septic Shock? *N Engl J Med*. 2018;378(9):860-861.
20. Rutegard M, Lagergren P, Johar A, et al. Time shift in early postoperative mortality after oesophagectomy for cancer. *Ann Surg Oncol*. 2015;22(9):3144-3149.
21. Talsma AK, Lingsma HF, Steyerberg EW, et al. The 30-day versus in-hospital and 90-day mortality after esophagectomy as indicators for quality of care. *Ann Surg*. 2014;260(2):267-273.
22. Kluge GH, Brinkman S, van Berkel G, et al. The association between ICU level of care and mortality in the Netherlands. *Intensive Care Med*. 2015;41(2):304-311.

23. Quraishi SA, Bittner EA, Blum L, et al. Prospective study of vitamin D status at initiation of care in critically ill surgical patients and risk of 90-day mortality. *Crit Care Med*. 2014;42(6):1365-1371.
24. Bukan RI, Moller AM, Henning MA, et al. Preadmission quality of life can predict mortality in intensive care unit--a prospective cohort study. *J Crit Care*. 2014;29(6):942-947.
25. Venkatesh B, Finfer S, Cohen J, et al. Adjunctive Glucocorticoid Therapy in Patients with Septic Shock. *N Engl J Med*. 2018;378(9):797-808.
26. Fizez T, Kerklaan D, Mesotten D, et al. Early versus Late Parenteral Nutrition in Critically Ill Children. *N Engl J Med*. 2016;374(12):1111-1122.
27. Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. *J Am Stat Assoc*. 1997;92(438):548-560.
28. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
29. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931.
30. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240(4857):1285-1293.
31. Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Stat Med*. 2014;33(14):2390-2407.
32. Finazzi S, Poole D, Luciani D, et al. Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLoS One*. 2011;6(2):e16110.
33. Steyerberg EW. *Clinical prediction models*. First Edition ed. New York, New York: Springer Science+Business Media; 2009.
34. Gemke RJ, van Vught J. Scoring systems in pediatric intensive care: PRISM III versus PIM. *Intensive Care Med*. 2002;28(2):204-207.

35. Shann F. The use and evaluation of pediatric index of mortality and pediatric risk of mortality in pediatric intensive care. *Pediatr Crit Care Med*. 2014;15(3):286-287.
36. Pollack MM, Holubkov R, Funai T, et al. The Pediatric Risk of Mortality Score: Update 2015. *Pediatr Crit Care Med*. 2016;17(1):2-9.

Figure legends

Figure 1: ROC curves for PICU mortality and 90-day mortality.

A. PICU mortality, AUC (95% CI): PIM3 0.894 (0.892-0.896), PRISM III 0.920 (0.918-0.921), $p=0.04$; B. 90-day mortality, AUC (95% CI): PIM3 0.867 (0.866-0.869), PRISM III 0.882 (0.880-0.884), $p=0.77$. The shaded regions represent the 95% CI band.

Abbreviations: AUC, area under the ROC curve; CI, confidence interval; PICU, pediatric intensive care unit; PIM, pediatric index of mortality; PRISM, pediatric risk of mortality; ROC, receiver-operating characteristics

Figure 2: Calibration belts for PICU mortality

The distribution of the patients stratified by endpoint (deaths above the x-axis, survivors below the x-axis) is indicated at the bottom of the graph. A. PIM3 calibration belt for PICU mortality: $p=0.58$; B. PRISM III calibration belt for PICU mortality: $p=0.04$. The shaded areas represent the 95% CI.

Abbreviations: CI, confidence interval; PIM, pediatric index of mortality; PRISM, pediatric risk of mortality

Figure 3: Calibration belts for 90-day mortality

The distribution of the patients stratified by endpoint (deaths above the x-axis, survivors below the x-axis) is indicated at the bottom of the graph. A. PIM3 calibration belt for 90-day mortality: $p=0.34$; B. PRISM III calibration belt for 90-day mortality: $p=0.04$. The shaded areas represent the 95% CI.

Abbreviations: CI, confidence interval; PIM, pediatric index of mortality; PRISM, pediatric risk of mortality