Comparison of the Automated Pediatric Logistic Organ Dysfunction-2 Versus Manual Pediatric Logistic Organ Dysfunction-2 Score for Critically Ill Children*

Michaël Sauthier, MD[1,2]; Florence Landry-Hould[2]; Stéphane Leteurtre, MD, PhD[3]; Atsushi Kawaguchi, MD, PhD[1,2]; Guillaume Emeriaud, MD, PhD[1,2]; Philippe Jouvet, MD, PhD[1,2]

Objectives: The Pediatric Logistic Organ Dysfunction-2 is a validated score that quantifies organ dysfunction severity and requires complex data collection that is time-consuming and subject to errors. We hypothesized that a computer algorithm that automatically collects and calculates the Pediatric Logistic Organ Dysfunction-2 (aPELOD-2) score would be valid, fast and at least as accurate as a manual approach (mPELOD-2).

Design: Retrospective cohort study.

Setting: Single center tertiary medical and surgical pediatric critical care unit (Sainte-Justine Hospital, Montreal, Canada).

Patients: Critically ill children participating in four clinical studies between January 2013 and August 2018, a period during which mPELOD-2 data were manually collected.

Interventions: None.

Measurements and Main Results: The aPELOD-2 was calculated for all consecutive admissions between 2013 and 2018 ($n = 5,279$) and had a good survival discrimination with an area under the receiver operating characteristic curve of 0.84 (95% CI, 0.81–0.88). We also collected data from four single-center studies in which mPELOD-2 was calculated ($n = 796$, 57% medical, 43% surgical) and compared these measurements to those of the aPELOD-2. For those patients, median age was 15 months (interquartile range, 3–73 mo), median ICU stay was 5 days (interquartile range, 3–9 d), mortality was 3.9% ($n = 28$). The intraclass correlation coefficient between mPELOD-2 and aPELOD-2 was 0.75 (95% CI, 0.73–0.77). The Bland-Altman showed a bias of 1.9 (95% CI, 1.7–2) and limits of agreement of −3.1 (95% CI, −3.4 to −2.8) to 6.8 (95% CI, 6.5–7.2). The highest agreement (Cohen's Kappa) of the Pediatric Logistic Organ Dysfunction-2 components was noted for lactate level (0.88), invasive ventilation (0.86), and creatinine level (0.82) and the lowest for the Glasgow Coma Scale (0.52). The proportion of patients with multiple organ dysfunction syndrome was higher for aPELOD-2 (78%) than mPELOD-2 (72%; $p = 0.002$). The aPELOD-2 had a better survival discrimination (area under the receiver operating characteristic curve, 0.81; 95% CI, 0.72–0.90) over mPELOD-2 (area under the receiver operating characteristic curve, 0.70; 95% CI, 0.59–0.82; $p = 0.01$).

Conclusions: We successfully created a freely available automatic algorithm to calculate the Pediatric Logistic Organ Dysfunction-2 score that is less labor intensive and has better survival discrimination than the manual calculation. Use of an automated system could greatly facilitate integration of the Pediatric Logistic Organ Dysfunction-2 score at the bedside and within clinical decision support systems. (*Pediatr Crit Care Med* 2020; 21:e160–e169)

Key Words: automatic data processing; children; clinical decision support systems; critical care; hospital mortality; organ dysfunction scores

Organ dysfunction assessment is central in critical care medicine (1). Even though mortality has substantially decreased in PICUs (2, 3) and is not the only outcome of interest, it is still the reference to build PICU severity scores (4–7). Although these scores are reliable and accurately reflect severity of illness, they are generally not used at the bedside because data collection is time-consuming and human error is very likely given the numerous elements of information required. These scores are frequently used as surrogate outcome measures in randomized clinical trials and are necessary to compare groups of patients. The Pediatric Logistic Organ Dysfunction-2 (PELOD-2) is a well-validated score based on 10 variables corresponding to five organ systems that is able to measure the severity of organ dysfunction (4, 8–11). Because all the required variables are stored in electronic medical records (EMRs), automation of data

*See also p. 397.

[1]Pediatric Intensive Care Unit, Department of Pediatrics, Sainte-Justine Hospital, Montreal, QC, Canada.

[2]Department of Pediatrics, Université de Montréal, Montreal, QC, Canada.

[3]Univ. Lille, CHU Lille, EA 2694 – Santé Publique: épidémiologie et qualité des soins, Service de réanimation pédiatrique, F-59000 Lille, France.

DOI: 10.1097/PCC.0000000000002235

collection would save time, provide perpetually updated information on patient clinical course and perhaps allow for estimation of therapeutic response. A few adult scores, specifically the Sequential Organ Failure Assessment (12–14) and Acute Physiology and Chronic Health Evaluation (APACHE) scores (15), have been automated with a performance that is comparable to that of manually collected scores. Even if EMRs have been used to calculate PELOD-2 in clinical studies (7, 10), no studies have compared manual calculation of the PELOD-2 to an automatic algorithm. The primary aim of this study was to validate an algorithm able to automatically calculate the PELOD-2 (aPELOD-2) based on the survival discrimination and the proportion of multiple organ dysfunction syndromes (MODSs). We hypothesized that aPELOD-2 has a survival discrimination that is similar to other PELOD-2 external validation studies. The secondary aim was to compare the pragmatic performance of the aPELOD-2 to a manually calculated PELOD-2 (mPELOD-2) measured in several research studies. Our hypothesis was that the performance of the aPELOD-2 is good and at least equivalent to that of the mPELOD-2.

MATERIALS AND METHODS

For the primary aim, we included data for all consecutive patients admitted to the PICU of Sainte-Justine University Hospital (Montreal, Canada) between January 8, 2013, and August 3, 2018. For the secondary aim, we included all patients that had a mPELOD-2 calculated for clinical studies undertaken at the PICU of Sainte-Justine University Hospital during the same period. For both the primary and secondary aims, we excluded patients 18 years old or older at admission. We followed the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) validation guidelines (16). The institutional review board approved this retrospective cohort study (reference number 2018-1587) and waived the need for individual consent.

Manually collected PELOD-2 scores came from four clinical studies: two were prospective studies (study one and three: ClinicalTrials.gov number NCT02613377 and NCT01977547, respectively) and two were retrospective studies with data collected by medical students (study two and four: manuscripts currently in preparation). The four studies comprised a broad good diversity of PICU patients: transfused patients, patients with respiratory failure, patients having undergone surgery for congenital heart disease, and patients with delirium.

Data Collection

Both mPELOD-2 and aPELOD-2 used the same EMR (IntelliSpace Critical Care and Anesthesia, Version F.01, Philips, Eindhoven, The Netherlands) as the data source. All PELOD-2 related fields were either directly recorded in the EMR (e.g., laboratory values and respiratory data) or typed in with an error checking mechanism that prevented physiologically incompatible values from being entered. Calculation of the Glasgow Coma Scale (GCS) was automatic and used drop-down menus to measure each function. The mPELOD-2 was collected by trained personnel that included medical students and research staff. Standard training for research clerks consisted of a cross validation of five to 10 subjects before they were allowed to collect data independently; all research staff were involved in multiple research studies requiring score calculations. Medical students had no experience with ICU scoring systems before they began data collection; their training consisted of basic education on the PELOD-2 score (as compared to other scores) and included hands on data acquisition while supervised to insure accuracy of data collection and full understanding of the data elements required.

Algorithm Description

The elements required to compute aPELOD-2 were collected using successive Structured Query Language queries (**Fig. 1**). The algorithm is freely available at https://github.com/sauthiem/aPELOD2 under the open-source GNU AGPL v3.0 license. The first step was to identify where the 10 variables required to compute the aPELOD-2 were stored in the database and import them into a temporary table structured on an entity-attribute-value (EAV) model that is robust to synonyms and can easily handle heterogeneous data (17, 18). To calculate the Pa_{O_2}/F_{IO_2} ratios, we extracted Pa_{O_2} values from the EAV table and looked for the last F_{IO_2} value available in the 60 minutes preceding Pa_{O_2} measurement. The Pa_{O_2}/F_{IO_2} ratio was then inserted as a new element in the EAV table. The 60 minutes maximum time lapse was based on local practice to calculate the Pa_{O_2}/F_{IO_2} ratio for the mPELOD-2 (step 2). For each data row, the day number was calculated on 24-hour intervals from the time of admission beginning with day one (step 3). Then, the algorithm pivoted the EAV table into a column-based structure, that is, one column per variable (step 4). During the pivoting process, data were agglomerated with the most abnormal value per patient and per day number. Finally, the most abnormal value was converted into points following the PELOD-2 pointing system and a summation was done to calculate the aPELOD-2 (step 5). As indicated in the original PELOD-2 methodology (4), missing values were considered normal (no point). Because studies 1, 3, and 4 used the calendar day to calculate the day number (from midnight to midnight), we adjusted the calculation of the aPELOD-2 to follow the same collection method as the mPELOD-2.

Statistical Analysis

We described the patient population using median and interquartile range (IQR) for continuous variables (age, ICU days, length of ventilation) and count with percentages for categorical variables and mortality. Statistical analysis was conducted in R 3.5.2 with the pROC package (19). We estimated survival discrimination with the area under the receiver operating characteristic curve (AUROC). The 95% CI and *p* value were calculated using the DeLong method (20). We compared the proportion of MODS present at admission; MODS was defined as the presence of two or more organs with one point or more (1).

We estimated the correlation between aPELOD-2 and mPELOD-2 using intraclass correlation coefficients (ICCs) with 95% CIs. We also calculated ICC among mPELOD-2

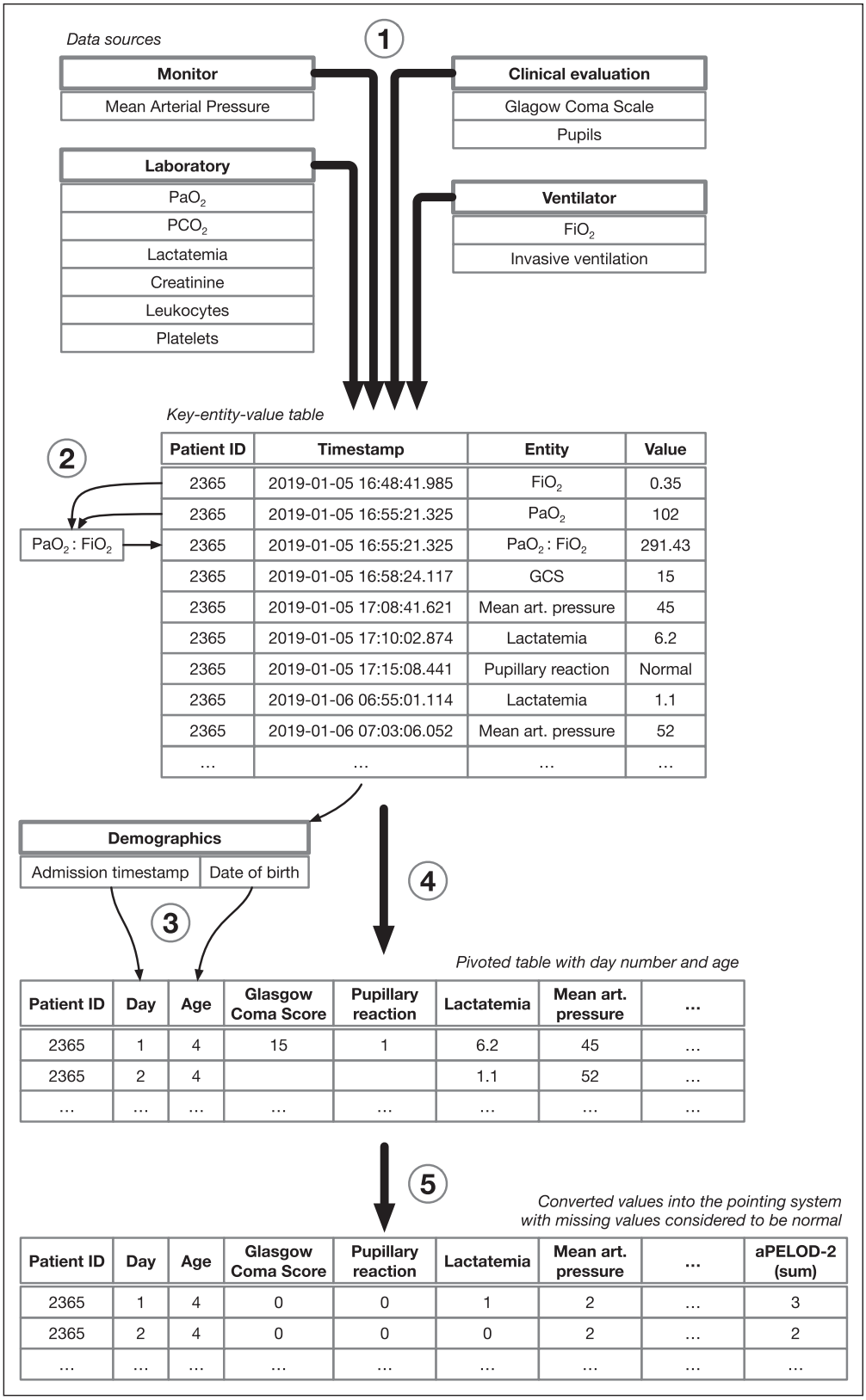

Figure 1. Automatically calculated Pediatric Logistic Organ Dysfunction-2 (aPELOD-2) algorithm data flow schema. Step 1: Data are collected from different sources and centralized into a key-entity-value structure. Step 2: Pao_2/Fio_2 calculation. Step 3: Age and day number calculation for each collected element. Step 4: Pivoting the keys into columns with selection of the most abnormal value per patient and per day. Step 5: The most abnormal value is converted into points and all categories are summed. GCS = Glasgow Coma Scale, ID = identifier.

scores when a patient was evaluated more than once. Because the overall correlation between aPELOD-2 and mPELOD-2 involved multiple "judges" (clinical studies in our case) and different evaluations (the mPELOD-2), we used a one-way random-effect model (ICC1,1) (21, 22). In all other cases, only two evaluations were compared (aPELOD-2 and mPELOD-2 or two mPELOD-2 studies); these methods were constant throughout all evaluations and a two-way random-effect model (ICC2,1) was used in those cases. The level of clinical significance of ICC was considered fair if between 0.4 and 0.59, good if between 0.6 and 0.74, and excellent if between 0.75 and 1 (23). We calculated inter-rater agreement for the different components of the PELOD-2 (categorical variables) with a linearly weighted Cohen's Kappa coefficient (24). Because Cohen's Kappa may not be reliable for rare observations or even impossible to calculate if agreement is perfect in a single category, we also reported overall and specific agreements for each component (25, 26). To illustrate specific agreements, we plotted the confusion matrix showing agreement proportion for each PELOD-2 variable between aPELOD-2 and mPELOD-2 (**Supplemental Fig. 1**, Supplemental Digital Content 1, http://links.lww.com/PCC/B173). Except in the case of a very low prevalence, Kappa agreement was interpreted as moderate if between 0.41 and 0.60, substantial if between 0.61 and 0.80, and almost perfect if between 0.81 and 1 (27). We also measured agreement between aPELOD-2 and mPELOD-2

using a Bland-Altman plot (28). Accuracy was estimated with bias (mean of the differences with 95% CIs) and precision was assessed with limits of agreement (± 1.96 × SDs of the differences with 95% CIs) and percentage error (29).

We compared aPELOD-2 and mPELOD-2 performance based on survival discrimination (AUROC). If more than one mPELOD-2 was collected for the same patient (i.e., the patient was included in two clinical studies), the average mPELOD-2 was used for comparison to the aPELOD-2. We compared aPELOD-2 and mPELOD-2 MODS estimation with a McNemar test. Statistical significance was defined as a *p* value of less than 0.05.

RESULTS

We included data from 5,279 patients admitted to PICU between January 8, 2013, and August 3, 2018 (**Table 1**). A total of 796 admission day mPELOD-2 calculations were collected in 725 children admitted between May 2013 and June 2018 who had been included in four different clinical studies. Median age was 15 months (IQR, 3–73 mo), female proportion was 46%, and overall mortality was 3.9%. The most frequent reasons for admission was elective surgery requiring postoperative care in PICU (40%), admission from the emergency department (28%), and admission from inpatient wards (22%).

aPELOD-2 Validation

The aPELOD-2 AUROC calculated on all consecutive admissions to PICU during the study period (*n* = 5,279 consecutive encounters, 2.9% mortality) was 0.84 (95% CI, 0.81–0.88) (**Fig. 2**). The proportion of MODS at admission was 62% (*n* = 3,250).

aPELOD-2 Comparison to mPELOD-2 From Four Clinical Trials

The median value for mPELOD-2 was 5 (IQR, 3–7) and for aPELOD-2 was 7 (IQR, 4–9.5) with a Pearson R^2 correlation coefficient of 0.68 (95% CI, 0.64–0.72; $p < 0.001$). Bland-Altman analysis (**Fig. 3**) showed a bias of +1.9 (95% CI, 1.7–2) for the aPELOD-2 over the mPELOD-2. The limits of agreements were calculated from –3.1 to 6.8 (95% CI, –3.4 to –2.8 and 6.5–7.2, respectively). The percentage error was 180%. Bland-Altman among pairs of manually calculated PELOD-2 (studies 1–3 and 3–4 with 29 and 25 patients, respectively) revealed a bias of 0.1 (95% CI, –0.4 to 0.6) and –0.4 (95% CI, –1.1 to 0.3), limits of agreements ± 2.6 and ± 3.3, and percentage error 144% and 99%, respectively. The ICC between aPELOD-2 and mPELOD-2 (**Table 2**) was 0.75 (95% CI, 0.73–77) with variability among studies: ICC values were 0.75, 0.85, 0.62, and 0.20 for studies 1, 2, 3, and 4, respectively. The ICC among pairs of mPELOD-2 varied from 0.58 to 0.92.

The weighted Cohen's Kappa coefficient was calculated for each PELOD-2 component for studies 2 and 3 (data were unavailable for studies 1 and 4) (**Table 3**). The highest coefficients were observed for lactate level (0.88), use of invasive ventilation (0.86), and creatinine level (0.82) while the lowest coefficients were noted for platelet count (0.64), leukocyte count (0.52), and the GCS (0.52). Because of the low prevalence of certain classes,

TABLE 1. Demographical Data

Variables	Validation on a PICU Database	Automatically Calculated Pediatric Logistic Organ Dysfunction-2 Comparison to Pediatric Logistic Organ Dysfunction-2 Scores Manually Collected in Clinical Trials				
		All Studies	Study 1	Study 2	Study 3	Study 4
Patients, *n* (%)	5,279 (100)	725 (100)	387 (49)[a]	171 (21)[a]	157 (20)[a]	81 (10)[a]
Mortality, *n* (%)	155 (2.9)	28[b] (3.9)	27 (7.0)	0 (0)	2 (1.3)	0 (0)
Females, *n* (%)	2,293 (43)	331 (46)	195 (50)	68 (40)	69 (44)	31 (38)
Age, mo, median (IQR)	30 (6–111)	15 (3–73)	19 (5–93)	26 (5–103)	12 (2–41)	0 (0–1)
ICU days, median (IQR)	2 (1–5)	5 (3–9)	6 (4–12)	4 (3–6)	5 (3–7)	7 (4–11)
Invasive ventilation days, median (IQR)	0 (0–1)	1 (0–4)	1 (0–5)	0 (0–1)	1 (1–4)	3 (1–7)
Admission origin, *n* (%)						
Planned surgery	1,477 (28)	293 (40)	116 (30)	46 (27)	117 (75)	69 (85)
Emergency department	2,066 (39)	204 (28)	118 (30)	70 (41)	21 (13)	0 (0)
Inpatient wards	1,230 (23)	170 (24)	114 (29)	38 (22)	18 (11)	9 (11)
Other hospitals	261 (5)	28 (4)	22 (6)	7 (4)	1 (1)	0 (0)
Unplanned surgery	206 (4)	23 (3)	13 (3)	7 (4)	0 (0)	3 (4)
Outpatient clinic	39 (1)	7 (1)	4 (1)	3 (2)	0 (0)	0 (0)

IQR = interquartile range.

[a]Proportion on the sum of the four studies (796 patients).

[b]One deceased patient was enrolled in study one and three.

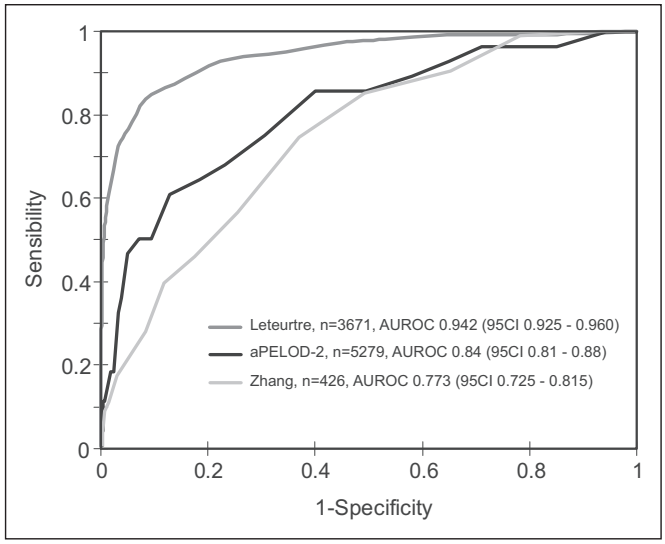

Figure 2. Comparison of the area under the receiver operating characteristic curve (AUROC) between the original Pediatric Logistic Organ Dysfunction-2 study (4), the automatically calculated Pediatric Logistic Organ Dysfunction-2 (aPELOD-2) and another external validation study by Zhang et al (11) with 95% CI.

there was a discrepancy between the overall agreement and the Kappa for leukocytes (agreement 98%, Kappa 0.53) and the PaO_2:FIO_2 ratio (agreement 95%, Kappa 0). The Kappa was impossible to calculate for pupillary reaction because there was a complete agreement on a single class (i.e., normal results only). Specific agreements (**Supplemental Table 1**, Supplemental Digital Content 1, http://links.lww.com/PCC/B173) were higher for near-normal results and decreased for more abnormal values. The confusion matrix plot (Supplemental Fig. 1, Supplemental Digital Content 1, http://links.lww.com/PCC/B173) showed a trend for the aPELOD-2 to overrate the severity of the organ dysfunction as compared with the mPELOD-2. Less values were reported as missing by the aPELOD-2 than the mPELOD-2 (**Supplemental Table 2**, Supplemental Digital Content 1, http://links.lww.com/PCC/B173).

MODS Screening

The proportion of patients (studies 2 and 3, $n = 328$) with two or more organ dysfunctions at admission day was higher ($p = 0.002$) when evaluated by the aPELOD-2 (78%, $n = 237$) than by the mPELOD-2 (72%, $n = 255$).

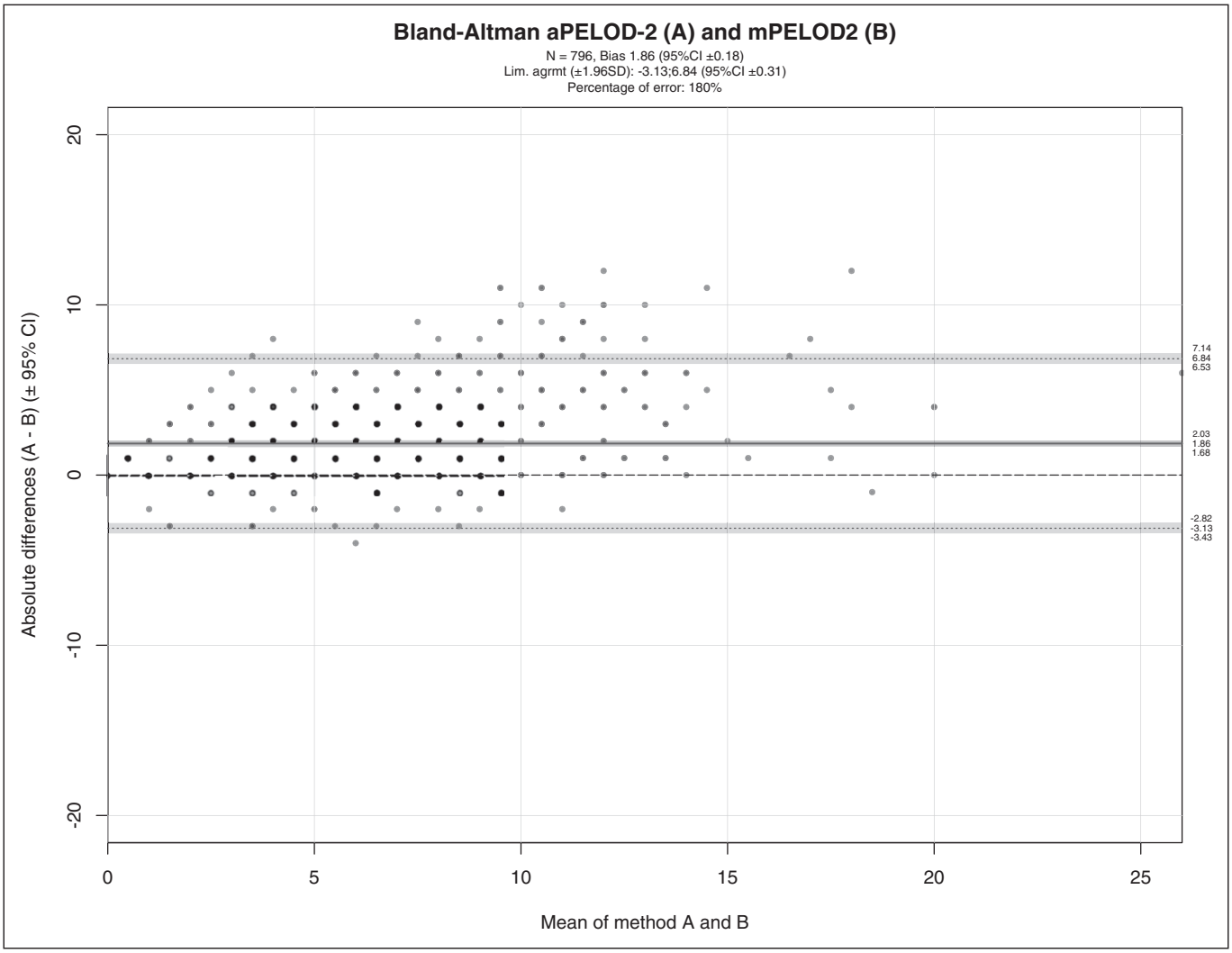

Figure 3. Bland-Altman plot between the automatically calculated Pediatric Logistic Organ Dysfunction-2 (aPELOD-2) (method A) and manually calculated Pediatric Logistic Organ Dysfunction-2 (mPELOD-2) (method B) with bias, limits of agreement and 95% CI.

TABLE 2. Intraclass Correlation Coefficients

Compared Groups	No. of Compared Scores	Intraclass Correlation Coefficients (95% CI)
aPELOD-2 vs mPELOD-2	796	0.75 (0.73–0.77)
aPELOD-2 vs study 1	387	0.75 (0.41–0.87)
aPELOD-2 vs study 2	171	0.85 (0.78–0.90)
aPELOD-2 vs study 3	157	0.62 (0.13–0.86)
aPELOD-2 vs study 4	80	0.20 (0–0.46)
mPELOD-2 study 1 vs 3	29	0.92 (0.85–0.96)
mPELOD-2 study 3 vs 4	25	0.71 (0.46–0.86)

aPELOD-2 = automatically calculated Pediatric Logistic Organ Dysfunction-2, mPELOD-2 = manually calculated Pediatric Logistic Organ Dysfunction-2.

aPELOD-2 and mPELOD-2 Survival Discrimination

Survival discrimination was similar between the aPELOD-2 (AUROC, 0.74; 95% CI, 0.63–0.85) and mPELOD-2 (AUROC, 0.70; 95% CI, 0.59–0.81) ($p = 0.15$) (**Fig. 4*A***). However, when the aPELOD-2 was calculated using the first 24 hours after the admission as recommended in the original PELOD-2 methodology (**Fig. 4*B***), the aPELOD-2 AUROC increased (0.81; 95% CI, 0.72–0.90) and became significantly higher than the mPELOD-2 ($p = 0.01$).

mPELOD-2 Computation

In order to explore the causes of disagreement between mPELOD-2 and aPELOD-2, we verified the manual calculation done for the only study that provided detailed data on manual collection (study 3). We found 10% disagreement between the indicated mPELOD-2 and the verified one; correction did not significantly improve the ICC with aPELOD-2 (0.65; 95% CI, 0.20–0.82 vs 0.62; 95% CI, 0.13–0.81).

Time Estimated to Calculate aPELOD-2

The algorithm was able to calculate a single aPELOD-2 score in approximatively 0.03 seconds.

DISCUSSION

The performance of the aPELOD-2 algorithm was good (AUROC 0.84) and was similar to AUROC values between 0.76 and 0.94 reported in the literature (Fig. 1) (7–11, 30, 31). Proportion of MODS (62%) was also similar to that reported in the literature (55%) (1). The intraclass correlation between aPELOD-2 and mPELOD-2 was between good and excellent, but with important variation among studies. Data collection was done by medical students for both the study with the lowest correlation (study 4) and the study with the highest correlation (study 2). Because both studies observed no mortality, it was impossible to use survival discrimination as a surrogate for the quality of data gathering. We interpret this as an example of inter-rater variability when nonprofessional raters collect data. In the literature, the inter-rater correlation has not been formally studied for the PELOD-2 score. However, the data collection process for the first version of the PELOD, which had an ICC of 0.79 and 0.86 on a subset of the original cohort, was very similar to that which occurred for PELOD-2 data collection in our study (32). The APACHE II score ICC has been evaluated in a dedicated prospective blinded study comparing three specifically trained raters (33) and had an excellent overall score (ICC 0.9); however, clinical interpretation for some components of the score was as low as 0.40. The Simplified Acute Physiologic Scores (SAPSs) II and III have also been evaluated with trained medical personnel; authors reported an overall ICC of 0.84 and 0.80 for SAPS II and III, respectively (34). This is congruent with our findings in which the ICC between professional research clerks (study 1 and 3) was excellent (0.92) and decreased when less experienced raters were involved. In this context, an ICC of 0.75 between aPELOD-2 and mPELOD-2 strengthens the validity of the aPELOD-2. Furthermore, the aPELOD-2 does not need any specific training to have perfect reproducibility. On the other hand, if the aPELOD-2 cannot be used, these data suggest there might be benefit in using highly qualified personnel with standardized training for mPELOD-2 data collection.

The aPELOD-2 algorithm significantly outperformed the manual PELOD-2 score. However, the superiority of the aPELOD-2 discrimination may be due to the calculation by 24-hour intervals starting at admission time. The only study that collected mPELOD-2 based on the 24-hour definition (study 2) had a nil mortality rate, preventing survival discrimination comparison.

Overall aPELOD-2 scores were slightly higher than mPELOD-2 (Supplemental Table 1; and Supplemental Fig. 1, Supplemental Digital Content 1, http://links.lww.com/PCC/B173). This can be explained, at least partially, by either a better sensitivity or a lower specificity. The algorithm may be more prone to include possible erroneous data because it cannot disregard abnormal data based on the clinical context. For example, laboratory results are involved in seven of the 10 components of the PELOD-2 score. Clinically suspected erroneous laboratory results are usually repeated or amended in clinical practice, but never erased from the EMR. Research assistants may decide during manual PELOD-2 scoring to keep or ignore laboratory results based on their clinical value. Thus, these laboratory results could be picked by the algorithm as the most abnormal value without consideration of whether the data are clinically valid. Regarding continuous data such as blood pressure, even EMRs with clinical validation may be subject to errors (35). To limit this risk, future scores could base their selection of abnormal continuous data on medians or percentiles, known to be simple and robust to minimize erroneous data, raw signal analysis with filtering (35) or machine learning algorithms (36). Furthermore, missing values are imputed as normal values for most severity scores including PELOD-2, but future scores may want to distinguish normal from missing values (Supplemental Table 2, Supplemental Digital Content 1, http://links.lww.com/PCC/B173). The original PELOD-2 definition did not account for the increased amount of data collected in modern ICUs such as that provided by continuous

TABLE 3. Cohen's Kappa and Overall Agreement on the 10 Components of the Pediatric Logistic Organ Dysfunction-2 Score

Pediatric Logistic Organ Dysfunction-2 Component	Study	Overall Agreement (%)	Cohen's Kappa (95% CI)
Glasgow Coma Score	2	96	0.84 (0.72–0.95)
	3	77	0.06 (–0.02 to 0.14)
	2+3	87	0.52 (0.39–0.64)
Pupillary reaction	2	100	—[a]
	3	100	—[a]
	2+3	100	—[a]
Lactatemia	2	98	0.88 (0.76–1)
	3	96	0.88 (0.79–0.98)
	2+3	97	0.88 (0.81–0.95)
Mean arterial pressure	2	84	0.77 (0.69–0.85)
	3	73	0.59 (0.49–0.7)
	2+3	78	0.68 (0.62–0.75)
Creatinine	2	88	0.76 (0.67–0.86)
	3	94	0.88 (0.81–0.96)
	2+3	91	0.82 (0.76–0.88)
Invasive ventilation	2	94	0.86 (0.78–0.94)
	3	93	0.78 (0.65–0.9)
	2+3	93	0.86 (0.81–0.92)
Pao_2: FIo_2	2	99	0[b]
	3	90	–0.02 (–0.05 to 0.01)
	2+3	95	–0.02 (–0.03 to 0)
Pco_2	2	92	0.75 (0.61–0.88)
	3	87	0.67 (0.55–0.79)
	2+3	90	0.71 (0.62–0.8)
Leukocytes	2	99	0.5 (–0.11 to 1)
	3	97	0.53 (0.17–0.89)
	2+3	98	0.52 (0.22–0.83)
Platelets	2	92	0.75 (0.64–0.86)
	3	74	0.53 (0.42–0.65)
	2+3	84	0.64 (0.56–0.72)

[a]Complete agreement with a null variance.
[b]Null variance in one group.

data stream from monitors (37) and the associated risk of having a single outlier that needs to be validated as clinically relevant data. On the other hand, the algorithm will not miss any abnormal result recorded in the EMR, regardless of the amount of data to analyze. As part of this study, we extracted approximatively 3.3 million values required by the 10 components of the PELOD-2 in 5,279 subjects in about 15 minutes. This amount of data is impossible to process for humans in a reasonable time period. Based on previous internal data, our institution estimates that about 20 minutes per day will be required for a research assistant to collect data for a daily PELOD score. Thus, if the PELOD-2 was systematically calculated on a daily basis in a center similar to ours (1,100 admissions per year, 6,700 patient-days per year), this could save 2,200 work hours (1.2 full time equivalent). Furthermore, all the other steps required by the PELOD-2 score (age and day calculation, Pao_2/FIo_2 calculation, normal values that depend on the patient age and summation of the different components) all comprise

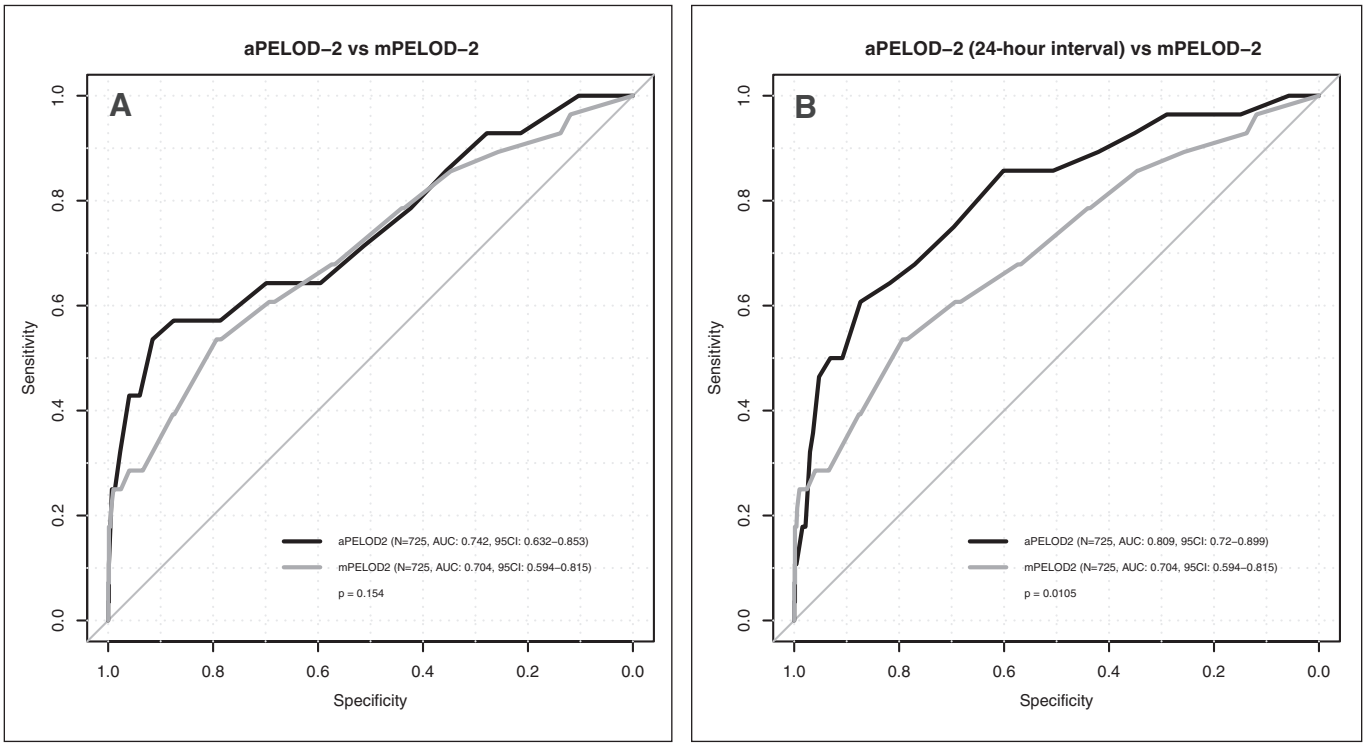

Figure 4. Death discrimination using the area under the receiver operating characteristic curve (AUC) for day-1 Pediatric Logistic Organ Dysfunction-2. Manually calculated Pediatric Logistic Organ Dysfunction-2 (mPELOD-2) curve is the same in both (**A**) and (**B**). **A**, Automatically calculated Pediatric Logistic Organ Dysfunction-2 (aPELOD-2) was calculated using data from the day of the admission (from 0:00 AM to 11:59 PM). **B**, aPELOD-2 was calculated using data from the first 24-hr after the admission.

a risk for error that a computer could easily avoid. For example, we found 10% of disagreement among mPELOD-2 calculation and also noticed that some mPELOD-2 scores reported three points in a category (platelets) for which the maximum is two (Supplemental Fig. 1, Supplemental Digital Content 1, http://links.lww.com/PCC/B173).

Components of severity scores that need clinical input are known to have a lower agreement between human raters or between algorithm and human (13, 14, 33, 38). Indeed, the lowest agreement in this study was for the GCS, possibly because only a human can ascertain whether the clinical context in which the GCS is measured truly represents the actual severity of the neurologic status and is not affected by sedation or the need for invasive ventilation.

The strengths of our study are the number of patients included in the survival discrimination analysis and the quality of the validation process. To our knowledge, this study is the largest that has compared an automatic calculation of a severity score to the manual equivalent in pediatric critical care. Furthermore, the open-source license makes the algorithm available for integration into constantly updated clinical decision support systems.

Our study has limitations. First, data come from specific clinical studies that may bring a selection bias. Nevertheless, the performance of the aPELOD-2 on all our consecutive encounters is very good and comparable to that in the literature. To minimize the possible impact of the latter limitation, we plan to conduct a similar study in several PICUs. Second, the percentage error in the Bland-Altman analysis is high (180%). This indicates that the limits of agreement are proportionally high compared to the mean value of the referenced method. An upper limit of 30% was suggested for adult cardiac output studies (39). In other contexts, this requires careful interpretation, especially when the variable is discrete and broad as seen in the PELOD-2 (28, 40). We compared this result to the percentage error among the different mPELOD-2 studies and found that they were high as well (99% and 144%). Therefore, the interpretation of the percentage error is limited in the aPELOD-2 validation process. Third, there is underrepresentation of mortality and patients with severe organ failure that certainly limits parts of the validation (such as the pupillary reaction). This as a consequence of global PICU mortality improvement; a large multicenter study would be required to address this limitation. Finally, the algorithm does not account for cyanotic status in children with congenital heart disease. Therefore, the Pao_2/Fio_2 ratio is calculated instead of being set to normal. Fortunately, the impact of this limitation is minimized by the PELOD-2 strict threshold of 60 which implies that all normally oxygenated cyanotic patient (Pao_2 40 mm Hg) with a Fio_2 below 67% would still be counted as normal. Future upgrades of the EMR will correct this limitation.

CONCLUSIONS

The aPELOD-2 provides a valid estimation of the PELOD-2 score that is fast, less labor intensive and well correlated with the mPELOD-2. Furthermore, the algorithm is freely available.

Use of the aPELOD-2 could occupy an important place within clinical decision support systems in pediatric critical care as well as serve for research purposes at the bedside. We have found that the aPELOD-2 has a better survival discrimination than the mPELOD-2 but recognize that some components, such as the GCS, may be better evaluated by the mPELOD-2. Our next steps will be to use the algorithm within a larger multicenter dataset in order to improve the algorithm on the component that requires clinical judgment and to increase robustness against erroneous data.

ACKNOWLEDGMENTS

We would like to thank Dr. Adrienne Randolph and Dr. Marisa Tucci for comments that greatly improved the article. We also thank Dr. Laurence Ducharme-Crevier, Dr. Geneviève Du Pont-Thibodeau, and Dr. Marisa Tucci, principal investigators of the different studies that made the automatically calculated Pediatric Logistic Organ Dysfunction-2 validation possible and the research staff at CHU Sainte-Justine for collecting the manually calculated Pediatric Logistic Organ Dysfunction-2 scores.

Drs. Sauthier, Landry-Hould, Leteurtre, Emeriaud, and Jouvet participated in the design of the automated Pediatric Logistic Organ Dysfunction-2 calculator and Dr. Sauthier wrote the algorithm. Data collection was performed by Drs. Landry-Hould and Sauthier. Data analysis, data interpretation, and article drafting was performed by Drs. Sauthier, Landry-Hould, Leteurtre, Emeriaud, and Jouvet. Statistical validation was performed by Dr. Sauthier. Statistical revision was performed by Dr. Kawaguchi. All authors participated in the critical revision and final approval of the article.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/pccmjournal).

This research and development study was supported by grants from the "Fonds de Recherche du Québec – Santé (FRQS)."

Dr. Sauthier received funding from the Fonds de recherche du Québec en Santé and the Université de Montréal Faculty of Medicine. Dr. Landry-Hould received funding from IVADO, programme excellence. Dr. Kawaguchi received funding from the Fonds de recherche du Québec en Santé and Women and Children's Health Research Institute. Dr. Emeriaud received funding from the Fonds de recherche du Québec en Santé (research scholarship). Dr. Jouvet received funding from Rona, Sainte-Justine Hospital; he has been an invited speaker for Air Liquide Santé; he received medical devices on loan from Philips Healthcare; and he received financial support from the public research agency of Quebec (Fonds de recherche Québec-Santé) and from the Quebec Ministry of Health. Dr. Leteurtre disclosed that she does not have any potential conflicts of interest.

For information regarding this article, E-mail: philippe.jouvet@umontreal.ca

---

REFERENCES

1. Leteurtre S, Duhamel A, Deken V, et al; Groupe Francophone de Réanimation et Urgences Pédiatriques: Daily estimation of the severity of organ dysfunctions in critically ill children by using the PELOD-2 score. *Crit Care* 2015; 19:324
2. 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; 11:549–555
3. Pinto NP, Rhinesmith EW, Kim TY, et al: Long-term function after pediatric critical illness: Results from the survivor outcomes study. *Pediatr Crit Care Med* 2017; 18:e122–e130
4. Leteurtre S, Duhamel A, Salleron J, et al; Groupe Francophone de Réanimation et d'Urgences Pédiatriques (GFRUP): PELOD-2: An update of the PEdiatric logistic organ dysfunction score. *Crit Care Med* 2013; 41:1761–1773
5. Pollack MM, Holubkov R, Funai T, et al; *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Collaborative Pediatric Critical Care Research Network: The pediatric risk of mortality score: Update 2015. *Pediatr Crit Care Med* 2016; 17:2–9
6. Straney L, Clements A, Parslow RC, et al; ANZICS Paediatric Study Group and the Paediatric Intensive Care Audit Network: Paediatric index of mortality 3: An updated model for predicting mortality in pediatric intensive care*. *Pediatr Crit Care Med* 2013; 14:673–681
7. Matics TJ, Sanchez-Pinto LN: Adaptation and validation of a pediatric sequential organ failure assessment score and evaluation of the sepsis-3 definitions in critically ill children. *JAMA Pediatr* 2017; 171:e172352
8. Gonçalves JP, Severo M, Rocha C, et al: Performance of PRISM III and PELOD-2 scores in a pediatric intensive care unit. *Eur J Pediatr* 2015; 174:1305–1310
9. El-Nawawy A, Mohsen AA, Abdel-Malik M, et al: Performance of the pediatric logistic organ dysfunction (PELOD) and (PELOD-2) scores in a pediatric intensive care unit of a developing country. *Eur J Pediatr* 2017; 176:849–855
10. Schlapbach LJ, Straney L, Bellomo R, et al: Prognostic accuracy of age-adapted SOFA, SIRS, PELOD-2, and qSOFA for in-hospital mortality among children with suspected infection admitted to the intensive care unit. *Intensive Care Med* 2018; 44:179–188
11. Zhang L, Huang H, Cheng Y, et al: [Predictive value of four pediatric scores of critical illness and mortality on evaluating mortality risk in pediatric critical patients]. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue* 2018; 30:51–56
12. Harrison AM, Yadav H, Pickering BW, et al: Validation of computerized automatic calculation of the sequential organ failure assessment score. *Crit Care Res Pract* 2013; 2013:975672
13. Aakre C, Franco PM, Ferreyra M, et al: Prospective validation of a near real-time EHR-integrated automated SOFA score calculator. *Int J Med Inform* 2017; 103:1–6
14. Huerta LE, Wanderer JP, Ehrenfeld JM, et al; SMART Investigators and the Pragmatic Critical Care Research Group: Validation of a sequential organ failure assessment score using electronic health record data. *J Med Syst* 2018; 42:199
15. Beck BF, Kaboli PJ, Perencevich EN, et al: An automated computerized critical illness severity scoring system derived from APACHE III: Modified APACHE. *J Crit Care* 2018; 48:237–242
16. Collins GS, Reitsma JB, Altman DG, et al: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med* 2015; 162:55–63
17. Nadkarni PM, Brandt C: Data extraction and ad hoc query of an entity-attribute-value database. *J Am Med Inform Assoc* 1998; 5:511–527
18. Murphy SN, Weber G, Mendis M, et al: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; 17:124–130
19. Robin X, Turck N, Hainard A, et al: pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12:77
20. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; 44:837–845
21. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979; 86:420–428
22. McGraw KO, Wong SP: Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996; 1:30–46
23. Cicchetti DV: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994; 6:284–290
24. Cohen J: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70:213–220

25. Cicchetti DV, Feinstein AR: High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990; 43:551–558
26. Hripcsak G, Heitjan DF: Measuring agreement in medical informatics reliability studies. *J Biomed Inform* 2002; 35:99–110
27. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20:37–46
28. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1:307–310
29. McLean AS, Needham A, Stewart D, et al: Estimation of cardiac output by noninvasive echocardiographic techniques in the critically ill subject. *Anaesth Intensive Care* 1997; 25:250–254
30. Karam O, Demaret P, Duhamel A, et al; PlasmaTV investigators: Performance of the PEdiatric Logistic Organ Dysfunction-2 score in critically ill children requiring plasma transfusions. *Ann Intensive Care* 2016; 6:98
31. Wong JJ, Hornik CP, Mok YH, et al: Performance of the paediatric index of mortality 3 and paediatric logistic organ dysfunction 2 scores in critically Ill children. *Ann Acad Med Singapore* 2018; 47:285–290
32. Leteurtre S, Martinot A, Duhamel A, et al: Validation of the paediatric logistic organ dysfunction (PELOD) score: Prospective, observational, multicentre study. *Lancet* 2003; 362:192–197
33. Kho ME, McDonald E, Stratford PW, et al: Interrater reliability of APACHE II scores for medical-surgical intensive care patients: A prospective blinded study. *Am J Crit Care* 2007; 16:378–383
34. Strand K, Strand LI, Flaatten H: The interrater reliability of SAPS II and SAPS 3. *Intensive Care Med* 2010; 36:850–853
35. Hug CW, Clifford GD, Reisner AT: Clinician blood pressure documentation of stable intensive care patients: An intelligent archiving agent has a higher association with future hypotension. *Crit Care Med* 2011; 39:1006–1014
36. Johnson AE, Ghassemi MM, Nemati S, et al: Machine learning and decision support in critical care. *Proc IEEE Inst Electr Electron Eng* 2016; 104:444–466
37. Brossier D, El Taani R, Sauthier M, et al: Creating a high-frequency electronic database in the PICU: The perpetual patient. *Pediatr Crit Care Med* 2018; 19:e189–e198
38. Arts DG, de Keizer NF, Vroom MB, et al: Reliability and accuracy of Sequential Organ Failure Assessment (SOFA) scoring. *Crit Care Med* 2005; 33:1988–1993
39. Critchley LA, Critchley JA: A meta-analysis of studies using bias and precision statistics to compare cardiac output measurement techniques. *J Clin Monit Comput* 1999; 15:85–91
40. Giavarina D: Understanding Bland Altman analysis. *Biochem Med (Zagreb)* 2015; 25:141–151