Can J Anesth/J Can Anesth (2020) 67:981–991
https://doi.org/10.1007/s12630-020-01686-5

REPORTS OF ORIGINAL INVESTIGATIONS

External validation of a prognostic model for intensive care unit mortality: a retrospective study using the Ontario Critical Care Information System

Validation externe d'un modèle pronostique de la mortalité à l'unité de soins intensifs : une étude rétrospective fondée sur le Système d'information sur les soins aux malades en phase critique de l'Ontario

Fran Priestap, MSc · Raymond Kao, MD, MPH, FRCPC · Claudio M. Martin, MSc, MD, FRCPC

Received: 12 September 2019 / Revised: 21 February 2020 / Accepted: 5 March 2020 / Published online: 7 May 2020

Abstract

Purpose *To externally validate an intensive care unit (ICU) mortality prediction model that was created using the Ontario Critical Care Information System (CCIS), which includes the Multiple Organ Dysfunction Score (MODS).*

Methods *We applied the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) recommendations to a prospective longitudinal cohort of patients discharged between 1 July 2015 and 31 December 31 2016 from 90 adult level-3 critical care units in Ontario. We used multivariable logistic regression with measures of discrimination, calibration-in-the-large, calibration slope, and flexible calibration plots to compare prediction model performance of the entire data set and for each ICU subtype.*

Results *Among 121,201 CCIS records with ICU mortality of 11.3%, the C-statistic for the validation data set was 0.805. The C-statistic ranged from 0.775 to 0.846 among the ICU subtypes. After intercept recalibration to adjust the baseline risk, the mean predicted risk of death matched actual ICU mortality. The calibration slope was close to 1 with all CCIS data and ICU subtypes of cardiovascular and community hospitals with low ventilation rates. Calibration slopes significantly less than 1 were found for ICUs in teaching hospitals and community hospitals with high ventilation rates whereas coronary care units had a calibration slope significantly higher than 1. Calibration plots revealed over-prediction in high risk groups to a varying degree across all cohorts.*

Conclusions *A risk prediction model primarily based on the MODS shows reproducibility and transportability after*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12630-020-01686-5) contains supplementary material, which is available to authorized users.

F. Priestap, MSc (✉)
London Health Sciences Centre – Victoria Hospital, 800 Commissioner's Rd E, London, ON, Canada N6A 5W9
e-mail: fran.priestap@lhsc.on.ca

R. Kao, MD, MPH, FRCPC
London Health Sciences Centre – Victoria Hospital, 800 Commissioner's Rd E, London, ON, Canada N6A 5W9

Division of Critical Care, Department of Medicine, Schulich School of Dentistry and Medicine, Western University, London, ON, Canada

C. M. Martin, MSc, MD, FRCPC
London Health Sciences Centre – Victoria Hospital, 800 Commissioner's Rd E, London, ON, Canada N6A 5W9

Division of Critical Care, Department of Medicine, Schulich School of Dentistry and Medicine, Western University, London, ON, Canada

Lawson Health Research Institute, London, ON, Canada

intercept recalibration. Risk adjusting models that use existing and feasible data collection can support performance measurement at the individual ICU level.

Résumé

Objectif *Nous souhaitions faire une validation externe d'un modèle de prédiction de la mortalité aux unités de soins intensifs (USI) créé en utilisant le Système d'information sur les soins aux malades en phase critique (SISMPC) de l'Ontario, qui comporte le Score de défaillance multisystémique (MODS).*

Méthode *Nous avons appliqué les recommandations de communication transparente d'un modèle de prédiction multivarié pour le pronostic ou le diagnostic individuel TRIPOD à une cohorte longitudinale prospective de patients. Ces patients devaient avoir reçu leur congé entre le 1er juillet 2015 et le 31 décembre 2016 de 90 unités de soins intensifs de niveau 3 pour adultes en Ontario. Nous avons utilisé une méthode de régression logistique multivariée accompagnée de mesures de discrimination, d'étalonnage global, de pentes d'étalonnage et de graphiques d'étalonnage afin de comparer la performance du modèle de prédiction pour l'ensemble des données dans son intégralité et pour chaque sous-type d'USI.*

Résultats *Parmi les 121 201 dossiers du SISMPC présentant une mortalité à l'USI de 11,3 %, la statistique C pour l'ensemble de données de validation était 0,805. La statistique C allait de 0,775 à 0,846 parmi les sous-types d'USI. Après réétalonnage de l'ordonnée afin d'ajuster le risque de base, le risque prédit moyen de décès correspondait à la mortalité réelle à l'USI. La pente d'étalonnage était proche de 1 pour toutes les données du SISMPC et tous les sous-types d'USI des hôpitaux cardiovasculaires et communautaires ayant de faibles taux de patients ventilés. Des pentes d'étalonnage significativement inférieures à 1 ont été observées pour les USI dans les hôpitaux universitaires et les hôpitaux communautaires ayant des taux de patients ventilés élevés, alors que les unités de soins coronariens présentaient une pente d'étalonnage significativement supérieure à 1. Les courbes d'étalonnage ont révélé une sur-prédiction dans les groupes à risque élevé à des degrés variables dans toutes les cohortes.*

Conclusion *Un modèle de prédiction du risque se fondant principalement sur le score MODS a montré sa reproductibilité et son applicabilité après réétalonnage de l'ordonnée. Les modèles d'ajustement du risque qui s'appuient sur des collectes de données existantes et réalisables peuvent aider à mesurer la performance au niveau de l'USI individuelle.*

Keywords intensive care unit (ICU) · mortality · prognostic model · external validation

Benchmarking can be used to identify opportunities for quality improvement.[1] Performance or benchmarks can be monitored over time within a single practice, or compared across different practices. These methods for performance measurement and improvement require careful interpretation of the results and awareness of limitations.[2] In complex systems, such as intensive care units (ICUs), it can be difficult to compare measures of quality since patients present with heterogeneous illnesses and varied disease severity. Methods have been proposed to account for this heterogeneity, most commonly regression techniques to risk-adjust the measure of interest.[3-5]

An ideal benchmarking system will use data that are readily available and simple to interpret.[6] Ontario is the most populous province in Canada. In 2007, the Critical Care Information System (CCIS) was implemented by the provincial health ministry as part of a strategy to improve the quality and efficiency of the critical care system.[7] The CCIS includes a measure of organ dysfunction on ICU admission (Multiple Organ Dysfunction Score [MODS])[8] and daily nursing workload measures (Nine Equivalents Nursing Manpower Use Score [NEMS])[9]; however, this data has not been used to perform risk-adjustment, likely because validated models for this purpose are lacking. The ability of MODS to predict mortality has been reported in small, single-centre studies from Canada, Finland, and other countries.[10,11] We used CCIS data from the two medical-surgical ICUs in our hospital to develop and internally validate a prediction model for ICU mortality.[12] None of these models have been externally validated.

External validation of a prediction model's performance is an important and necessary process prior to clinical implementation.[13-16] Access to "big data" is increasing as evident by analysis of registry databases that contain electronic health records for thousands or even millions of patients from multiple practices and hospitals.[17] The CCIS is an example of a large e-health database that includes data from different types of ICUs, and thus provides an opportunity to assess both reproducibility (similar case-mix) and transportability (different but related populations) within the same study.[18] The objective of this study was to conduct and report a methodologically sound external validation using guidelines and referenced statistical articles from the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) explanation and elaboration document.[19]

Methods

Approval for this study was granted by the Western University Research Ethics Board on 15 February 2017. Requirement for consent was waived.

Study design

We used an independent population-based cohort to perform a validation study on a previously published ICU mortality prediction model.[12]

Data source

We used data from the Ontario CCIS for this study. The CCIS is a web-based data application that uses a combination of methods to capture data. Demographic data can be auto-populated directly from the hospital electronic admission, discharge, and transfer system, but most of the data are manually entered by clerical and clinical staff as appropriate. Data elements used in this study, a subset of those captured in CCIS, are shown in Electronic Supplementary Material (ESM), eTable 1. All ICUs in Ontario are required to enter data into the CCIS for all admissions.

Data were obtained for all level-3 ICU admissions between July 1 2015 and December 31 2016. Level 3 ICUs are defined as those providing life support and mechanical ventilation for more than 48 hours. Critical Care Services Ontario has organized the ICUs into groups based on ICU subtype (Table 1). The eligibility criteria, conditions, definitions, and measurements in this validation study were identical to those used in the original development study.

The minimum effective sample size for external validation has been reported as 100 outcome events.[20] The data set included over 13,500 deaths. All ICU subtype groups had well over 100 deaths except burn ICUs, which were excluded from the subgroup analyses.

The validation data set was first subject to administrative cleaning. We excluded admissions to pediatric and labour and delivery level-3 ICUs. Also excluded were records where patient age was reported as < 18 yr or > 115 yr, length of ICU stay was reported as 0 days (entry errors), or where duplicate MODS and/or NEMS entries were reported. For duplicate records, the record with the later time stamp was selected for linkage with the admission and discharge data. Finally, any records with missing predictor data were omitted from the analyses.

Complete case analyses were used to assess model performance. Records with missing data represented approximately 5% of all cases and exclusion of these cases was not considered a threat to the validity of the results.[21] The outcome of interest was ICU mortality. Predictor variables, available within the first 24 hr of critical care admission, were defined as follows: 1) age group (18–39, 40–79, $\geq$ 80 yr); 2) sex (M or F); 3) NEMS group (0–22, 23–29, $\geq$ 30); 4) MODS group (0, 1–4, 5–8, 9–12, $\geq$ 13); 5) admission source (operating room/postanesthesia care unit, emergency department, unit/ward, other hospital and other); 6) admitting diagnosis (cardiovascular/cardiac/vascular, respiratory, gastrointestinal, neurologic, trauma, other); and 7) readmission to critical care during the same hospital stay.[12]

Since we chose to restrict our analyses to variables contained within the CCIS data set, we modified our previously published model [12] by excluding the Charlson Comorbidity Index. eTable 1 (available as ESM) is

Table 1 CCIS level-3 ICU subtype groups and number of critical care units

Criteria	# Critical care units
Teaching hospitals (medical surgical ICU)	17
Community hospitals (medical surgical ICU) with ventilator patient day rate above the mean rate*	28
Community hospitals (medical surgical ICU) with ventilator patient day rate equal to or less than the mean rate*	23
Cardiac/cardiovascular unit	10
Coronary care units†	10
Burn units	2

*Ventilator patient day rate = (ventilator days/patient care days) * 100 based on fiscal year 2016–2017; mean rate = 43.61%

† Coronary care units that provide invasive ventilation for longer than 48 hr

CCIS = Critical Care Information System; ICU = intensive care unit

provided as supplemental digital content and shows the equation for Logit [ICU Mortality].

Statistical analyses

The relatedness of the development and validation data sets was reviewed using two approaches. First, the distribution of context-important patient characteristics, including predictors and outcomes, were compared. Descriptive analyses of these characteristics were performed for the development and validation data sets and for the latter, also stratified by CCIS ICU subtype. Continuous data elements are expressed as mean (standard deviation [SD]) or median [interquartile range (IQR)] as appropriate. Categorical data elements are reported as proportions. To quantify the extent of the relatedness in case-mix between the development and validation samples, a binary logistic regression model (membership model) was created to predict the probability that an individual record belonged to either sample.[22] Independent variables were the predictors and outcome from the prediction model. The discriminative ability of the model was quantified using its C-statistic with lower values indicating similarity between the data sets.

Three measures were used to assess the performance of the model in the validation data set: 1) calibration-in-the-large, 2) calibration slope, and 3) discrimination. Calibration-in-the-large represents the level of agreement between observed and predicted mortality. It was calculated as the logistic regression model intercept given that the calibration slope equals 1 ($\text{logit}(y)=a + \text{logit}(\hat{y})$).[22,23] Where calibration-in-the-large was significantly different from 0, intercept recalibration was performed by fitting a new logistic regression model with an intercept only and an offset term for the linear predictor. Calibration slope reflects whether predicted risks are appropriately scaled with respect to each other over the entire range of possible values. It was estimated from the recalibration model equation $\text{logit}(y)=a + b_{\text{overall}} \text{logit}(\hat{y})$.[22,24] Loess-based calibration plots were created with predicted risk on the x-axis and observed mortality on the y-axis to illustrate the agreement across the range of predicted risks.[23] Discrimination refers to the ability of the prediction model to separate individuals that died and those that survived. The concordance statistic was used to evaluate the discriminative value of the prediction model.

For those observations excluded from the analyses because of missing predictors, comparisons with the observations used in the validation were also made. All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC, USA).

Results

After applying the exclusion criteria, 121,201 records were available for external validation (Fig. 1). The demographic and clinical characteristics (predictors) and ICU mortality of the patient population included in the development model and external validation data set are shown in Table 2. The C-statistic for the membership model comparing the development data set to the entire CCIS cohort was 0.764. Values between 0.7 and 0.8 are generally considered to reflect acceptable discrimination[25] and in the case of this membership model, represent a data set that is somewhat related to the development data set, but not strongly so where a C-statistic of < 0.7 would be expected.

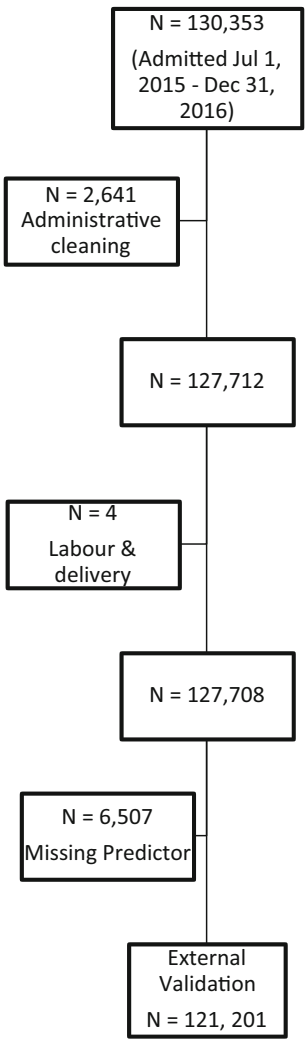

Fig. 1 Flow chart of patient records included in the external validation. Administrative cleaning includes the following: n = 1,609 (duplicates), $n = 427$ (admitted in error), $n = 88$ (ICU LOS = 0), $n = 511$ (age < 18 yr), $n = 6$ (age > 105 yr). ICU = intensive care unit; LOS = length of stay

Table 2 Baseline and clinical characteristics and outcomes of patients in the development and external validation data sets

	Development	External validation	
		Included	Missing
Total number of subjects, *n*	4,321	121,201	6,507
Sex			
% Female	42.7	40.2	38.7
N missing			3
Age (yr)			
0–39	13.5	9.1	8.7
40–79	72.8	73.4	72.5
≥ 80	13.7	17.5	18.8
N missing			0
ICU admission source			
Operating room/postanesthesia care unit	21.6	28.3	25.6
Other hospital	18.3	11.4	13.5
Emergency department	29.3	36.1	32.1
Other source*	8.8	9.7	15.0
Unit/ward	22.2	14.5	13.8
N missing			28
ICU admission diagnosis			
Cardiovascular/cardiac/vascular	15.1	43.1	55.6
Other diagnosis†	23.1	21.1	18.1
Gastrointestinal	10.5	6.1	5.7
Respiratory	32.5	19.3	13.7
Trauma	6.6	2.3	1.5
Neurologic	12.2	8.0	5.5
N missing			0
Multiple Organ Dysfunction Score (MODS)			
0	5.7	16.9	21.1
1–4	39.9	46.0	47.7
5–8	40.1	29.0	22.8
9–12	12.3	7.2	7.3
> 13	2.0	0.9	1.1
N missing			5607
Nine Equivalents Nursing Manpower Use Score (NEMS)			
0–22	12.9	36.4	43.1
23–29	32.5	26.3	23.6
≥ 30	54.6	37.3	33.3
N missing			2280
Re-admission to ICU (same hospital admission)	9.1	5.8	6.4
N missing			128
Mortality	22.8	11.2	11.9
N missing			28

† Other diagnosis includes patients with the following diseases: Metabolic/endocrine, Genitourinary, Musculoskeletal, Skin, Oncology, Hematology, Other

*Other source includes patients admitted from the following locations: Home – within or outside LHIN, Level 2 unit or step-down unit, Level 3 unit (medical/surgical or specialty unit), Complex continuing-care facility, Rehabilitation facility, Outside province, Other

ICU = intensive care unit; LHIN = Local Health Integration Network

This is confirmed by some key differences illustrated in Table 2. Specifically, the development population was younger, had a different source distribution (less from the operating room and emergency department, more from the ward and referrals from other hospitals), as well as higher levels of organ dysfunction upon admission, daily nurse workload, readmission, and ICU mortality. Admitting diagnosis also differed between the data sets with the development sample having a higher proportion of admissions for respiratory issues and a lesser proportion of cardiovascular-related admissions.

These same analyses were performed for each ICU subtype group. The discrimination of the membership models indicated varying degrees of relatedness to the development sample. Relatedness to the development sample was found in teaching hospital medical-surgical units (C-statistic = 0.660) and community hospital medical-surgical units with high rates of mechanical ventilation (C-statistic = 0.740) but discordance in community hospital medical-surgical units with low rates of mechanical ventilation (C-statistic = 0.836), cardiac/cardiovascular units (C-statistic = 0.969), and coronary care units (C-statistic = 0.974). eTable 2 (available as ESM) is provided as supplemental digital content and shows the characteristics and outcomes for each individual ICU subtype group compared with those for the entire cohort. The demographic and clinical profile of cases excluded from the analyses because of missing data were similar to those included in the external validation (Table 2), and as such, data were considered to be missing completely at random.

Calibration-in-the-large represents overall calibration of the model. Perfect agreement between observed and predicted values has an intercept value of 0. For all data combined and also for all ICU subtype groups except medical-surgical units in teaching hospitals, the intercept value was less than 0 indicating that the model over-predicted ICU mortality.[22] This over-estimation was greatest in cardiac/cardiovascular and coronary care units. In the medical-surgical units in teaching hospitals, the intercept value was greater than 0 showing a slight under-estimation of mortality (Table 3). Given the differences between actual ICU mortality and predicted risk, an intercept recalibration was performed for all models resulting in calibration-in-the-large values that are essentially 0.

The calibration plots in Figs 2a and 2b show that some over-prediction remains following intercept recalibration, specifically when the risk of death is higher. The extent of over-prediction varies across ICU subtype groups but represents a small proportion of patients.

The calibration slope reflects whether the predicted risks are scaled appropriately to each other over the complete range of predicted probabilities and was another measure used to evaluate the model's predictive performance in the validation samples. Calibration slopes not significantly different from 1 include all CCIS data, as well as community hospital medical-surgical units and cardiac/cardiovascular units. The calibration slope for teaching hospital medical-surgical units were significantly less than 1, showing higher variation in predicted probabilities (Table 3). Specifically, the variation between predicted and observed risks is too low for low-outcome risks and too high for high-outcome risks. The coronary care unit data set has a calibration slope significantly above 1 indicating too little variation in the predicted risks; predicted risks are systemically too high.

Discrimination for all CCIS data and the individual ICU subtype groups ranged from acceptable to very good (Table 3). The validation data sets with the lowest area under the curve (AUC) [IQR] were teaching hospital medical-surgical units (C = 0.781 [0.774 -0.788]) and cardiovascular/cardiac units C = 0.768 [0.747 - 0.789]). The data sets including all CCIS data and all other ICU subtype groups had areas under the curve greater than 0.80.

Discussion

We used a prospectively collected, population-based cohort to perform external validation on a risk prediction model for ICU mortality. We found that an intercept update was required, which greatly improved the calibration-in-the-large for the entire cohort as well as for all ICU subtype groups. Over-estimation for higher predicted risk groups remains, but this population represents relatively few patients. Since the intention of the model is for performance measurement and not individual patient prognosis, the model fit is acceptable for the entire cohort of ICUs.

The development and application of robust prognostic models are essential for valid performance measurement and many existing prognostic models have a limited life span because of changes in clinical practice and healthcare over time that can alter the risk of mortality for a given clinical situation. Prognostic models require periodic updating. Current prognostic models for mortality were published between 2005 and 2007 including Acute Physiology and Chronic Health Evaluation (APACHE) IV (AUC = 0.88),[5] Simplified Acute Physiology Score (AUC = 0.848,)[26] and Mortality Probability Admission Model (MPM0)-III (AUC = 0.823).[27] The organ dysfunction scores that assess the presence and severity of organ dysfunction include MODS (AUC = 0.695), Sequential Organ Failure Assessment (SOFA) (AUC = 0.776), and Logistic Organ Dysfunction Score (AUC =

Table 3 Predicted risk and model performance statistics for external validation of the entire CCIS cohort and for ICU subtype groups

	n	Observed ICU mortality		Predicted risk of death in ICU before intercept recalibration		Predicted risk of death in ICU after intercept recalibration		Calibration in the large before intercept recalibration (95% CI)	Calibration in the large after intercept recalibration (95% CI)	Calibration slope (βrisk)	Discrimination (C-statistic) (95% CI)
		(*n*)	%	Mean (SD)	Median [IQR]	Mean (SD)	Median [IQR]				
All Level 3 units	121,201	13,594	11.2	14.4	8.7	11.2	6.3	− 0.343	0.001	1.019	0.807
				(14.7)	[3.9–18.7]	(12.5)	[2.8–14.0]	(− 0.362 to − 0.324)	(− 0.018 to 0.021)	(1.001 to 1.037)	(0.804 to 0.811)
Teaching hospitals – medical/surgical units	28,894	4711	16.3	15.5	10.4	16.3	10.8	0.070	− 0.000	0.909	0.781
				(15.2)	[4.2–20.8]	(15.7)	[4.6–22.2]	(0.036 to 0.105)	(− 0.035 to 0.034)	(0.878 to 0.941)	(0.774 to 0.788)
Community hospitals – medical/surgical units with high ventilator patient day rate*	33,653	5398	16.0	16.4	10.0	16.1	9.6	− 0.037	− 0.006	0.982	0.816
				(17.1)	[3.7–22.3]	(16.9)	[3.6–21.6]	(− 0.069 to − 0.004)	(− 0.038 to 0.027)	(0.953 to 1.011)	(0.810 to 0.822)
Community hospitals – medical/surgical units with low ventilator patient day rate*	23,597	2100	8.9	11.6	7.5	9.1	5.5	− 0.350	− 0.032	1.048	0.810
				(12.7)	[3.5–16.3]	(10.7)	[2.5–12.5]	(− 0.399 to − 0.302)	(− 0.080 to 0.016)	(1.003 to 1.093)	(0.801 to 0.819)
Cardiac /cardiovascular units	18,929	608	3.2	14.6	11.4	3.0	1.9	− 1.799	0.064	1.062	0.768
				(12.4)	[7.3–16.3]	(3.8)	[1.3–3.0]	(− 1.883 to − 1.716)	(− 0.019 to 0.147)	(0.980 to 1.144)	(0.747 to 0.789)
Coronary care units	15,159	746	4.9	12.4	7.7	5.0	2.6	− 1.148	− 0.011	1.232	0.850
				(12.8)	[3.5–17.1]	(6.6)	[1.2–6.3]	(− 1.225 to − 1.070)	(− 0.088 to 0.067)	(1.156 to 1.307)	(0.836 to 0.863)

*Ventilator patient day rate = (ventilator days / patient care days) *100; based on fiscal year 2016–2017; mean rate = 43.61%; β = standardized regression (beta) coefficient; CCIS = Critical Care Information System; CI = confidence interval; ICU = intensive care unit; IQR = interquartile range; SD = standard deviation

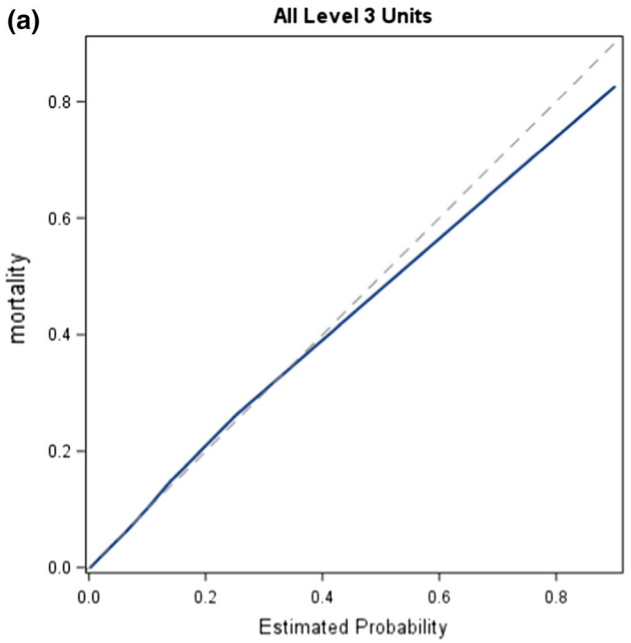

Fig. 2 a Loess-based calibration plots for validation of entire CCIS cohort. CCIS = Critical Care Information System. b Loess-based calibration plots for validation of individual ICU subtype groups. ICU = intensive care unit; TH = teaching hospitals; CH = community hospitals

0.805).[11] The AUC we report here for the entire cohort and for ICU subtype groups compares favourably with these other models.

The development model showed strong agreement between observed and expected mortality as assessed using the Hosmer-Lemeshow goodness-of-fit test. Limitations of this decile-based analysis include the influence of sample size and the arbitrary selection of the risk categories.[28-30] In this external validation, calibration was assessed using loess-based calibration plots, calibration-in-the-large, and calibration slope.[23] Although the results are not directly comparable, the underlying conclusions are that the model has acceptable calibration in both the development and validation data sets, indicating good overall agreement between observed and expected ICU mortality.

Discriminative ability increased slightly in this external validation and the membership model did indicate some case-mix differences. We anticipated that a data set containing over 120,000 patients would include a more diverse case-mix than the developmental model. Differences in case-mix can include the distribution of predictor values, varied participant or setting characteristics, and incidence of the outcome.[18] This increase in heterogeneity would enhance discriminative ability in the validation cohort, and has several effects on model performance across different settings and populations.[31,32] In fact, case-mix variation can lead to differences in the performance of a prediction model, even when the true predictors' effects are consistent.[31]

Benchmarking is an approach to identify and implement best practices.[1,33] Indicators selected for benchmarking can be compared over time within a single unit or practice, across units or practices or against a predetermined goal. Many potential indicators will not require risk or case-mix adjustment, while this will be needed for most patient-related outcomes such as mortality and length of stay. We caution against use of simple rank ordering or comparisons of one unit to another since regression models, such as the one we report, provide an estimated risk based on the average of the entire cohort. While our recalibration has reduced the bias across this cohort, estimates for subgroups or individual ICUs will remain biased. As can be seen in our data, it appears that teaching hospitals perform worse than average, community hospitals with high ventilator usage perform better than average, and cardiac units perform much better than average. Nevertheless, this would be a false conclusion since the differences across subgroups must cancel out across the entire cohort. At most, evaluation of subgroups or individual ICU results should only be compared with the average estimated performance and include confidence intervals.[3] Models could be recalibrated for specific ICU subtypes but this involves subjective categorization of units and will not resolve the bias for individual ICUs. One randomized trial used quantiles to identify achievable performance levels for groups of units and reported improved performance in individual units.[34] Ultimately, we believe that models such as these should be used to monitor performance over time only within individual ICUs. One such approach incorporates risk-adjusted measures into statistical process control methods.[35,36]

There are numerous strengths to this study. First, the breadth of the units that submit data to the CCIS allows for testing of both reproducibility (similar ICU subtype groups) and transportability (different ICU subtype groups), and the size of the CCIS data set provided ample statistical power for the required analyses. The TRIPOD framework indicates that a model's predictive performance should be evaluated in relation to subgroups of interest, such as age or sex, specific settings or population rather than just across all individuals combined, which can mask any deficiencies in the model.[19] It is increasingly recognized that the predictive performance of a model tends to vary across settings, populations, and periods,[22,31,37,38] which implies there is often heterogeneity in model performance and that multiple external validation studies are needed to fully appreciate the generalizability of a prediction model.[22] In this study, we have conducted subgroup analyses for each ICU

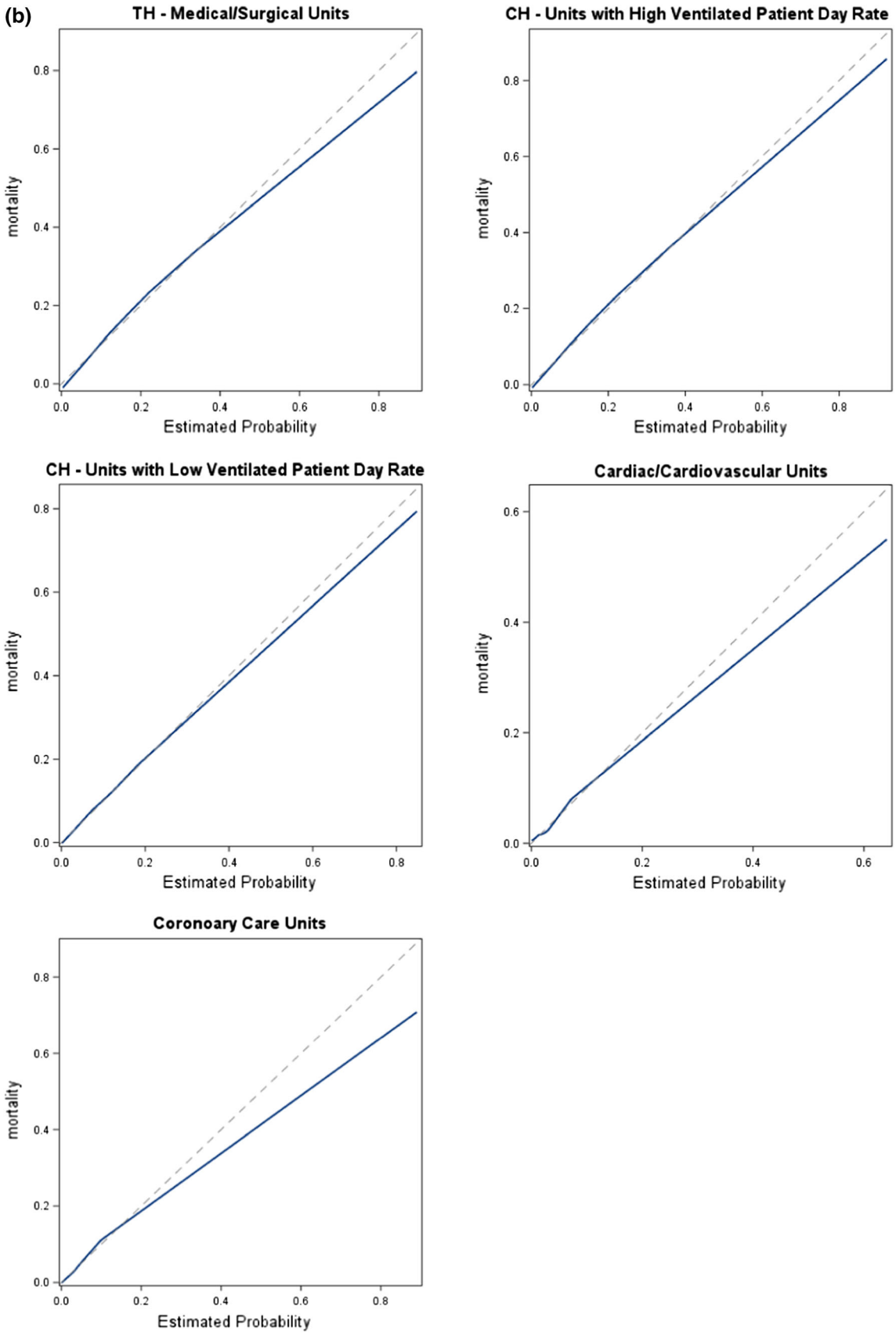

Fig. 2 continued

subtype to evaluate performance in specific ICU patient populations. Another strength is adherence to the TRIPOD guidelines, which include references to appropriate analytic methods and complete reporting of the results.[15,22,39,40] Next, both MODS and NEMS are relatively easy to collect, making this prediction tool more apt for risk-adjustment compared with more complex scoring systems. MODS requires only eight routinely collected variables and, in contrast to SOFA, is not dependent on treatment.[41] NEMS assesses ICU resource utilization and efficiency that has been validated as a nurse workload measure in large cohorts of ICU patients.[42] It is easy to use with minimum inter-observer variability,[9,42] but has not been evaluated as a mortality or risk prediction tool.

Limitations of this study include our inability to adjust for chronic health status as these data are not captured in the CCIS. Linkage to other data sets containing comorbidity data such as the Canadian Institute for Health Informatics Discharge Abstract Database could resolve this limitation, but we did not have access to identifiable patient information and such linkage was not possible. Another limitation is that, although ICU mortality is a proximal metric that can be used to evaluate quality of care in the ICU and ultimately improve patient outcomes, ICU survival is not a patient-centred goal. We found a low frequency of patients within the range of severity where mortality is over-predicted; however, this would need to be monitored regularly to ensure that results are interpreted correctly. Also, we could not evaluate the burn ICU subtype group accurately because of the low number of deaths. Finally, although there are no published studies on the accuracy of the CCIS data, we previously reported that inter-observer variability in data collection appears to be randomly distributed.[43]

Conclusion

Following an intercept update to adjust for the difference in mortality between the development and validation data sets, our ICU mortality prediction model performs well and shows both reproducibility and transportability. Some ICU subtype groups show inferior model fit compared with others, but the over-estimation of mortality occurs primarily in risk groups with low prevalence and thus has a minimal impact on overall calibration. These models could be used to provide risk-adjusted mortality rates to support performance measurement over time within individual ICUs using data that is easy and feasible to collect. Since the model represents an average of all the patients in the cohort, we recommend it should not be used for simple comparisons between ICUs or ICU subtypes.

Author contributions *Fran Priestap* contributed substantially to all aspects of the manuscript, including study conception and design; acquisition, analysis and interpretation of data; and drafting the article. *Raymond Kao* contributed substantially to study conception and design; acquisition, analysis, and interpretation of data; and drafting the article. *Claudio Martin* contributed substantially to study conception and design; acquisition, analysis and interpretation of data; and drafting the article.

Conflicts of interest None.

Financial statement Academic Medical Organization of Southwestern Ontario (AMOSO) grant

Editorial responsibility This submission was handled by Dr. Sangeeta Mehta, Associate Editor, *Canadian Journal of Anesthesia.*

References

1. *Keenan SP*, *Martin CM*, *Kossuth JD*, *Eberhard J*, *Sibbald WJ*. The Critical Care Research Network: a partnership in community-based research and research transfer. J Eval Clin Pract 2000; 6: 15-22.
2. *Lilford R*, *Mohammed MA*, *Spiegelhalter D*, *Thomson R*. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. Lancet 2004; 363: 1147-54.
3. *Breslow MJ*, *Badawi O*. Severity scoring in the critically ill: part 1–interpretation and accuracy of outcome prediction scoring systems. Chest 2012; 141: 245-52.
4. *Breslow MJ*, *Badawi O*. Severity scoring in the critically ill: part 2: maximizing value from outcome prediction scoring systems. Chest 2012; 141: 518-27.
5. *Zimmerman JE*, *Kramer AA*, *McNair DS*, *Malila FM*. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med 2006; 34: 1297-310.
6. *Pronovost PJ*, *Berenholtz SM*, *Ngo K*, *McDowell M*, *et al*. Developing and pilot testing quality indicators in the intensive care unit. J Crit Care 2003; 18: 145-55.
7. *Bell R, Robinson L.* Final Report of the Ontario Critical Care Steering Committee - March 2005. Available from URL: https://www.criticalcareontario.ca/EN/Toolbox/Overview%20of%20Ontarios%20Critical%20Care%20System/Final%20Report%20of%20the%20Ontario%20Critical%20Care%20Steering%20Committee%20(2005).pdf (accessed March 2020).
8. *Marshall JC*, *Cook DJ*, *Christou NV*, *Bernard GR*, *Sprung CL*, *Sibbald WJ*. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Crit Care Med 1995; 23: 1638-52.
9. *Reis MD*, *Moreno R*, *Iapichino G*. Nine equivalents of nursing manpower use score (NEMS). Intensive Care Med 1997; 23: 760-5.
10. *Zygun DA*, *Laupland KB*, *Fick GH*, *Sandham JD*, *Doig CJ*. Limited ability of SOFA and MOD scores to discriminate outcome: a prospective evaluation in 1,436 patients. Can J Anesth 2005; 52: 302-8.
11. *Pettila V*, *Pettila M*, *Sarna S*, *Voutilainen P*, *Takkunen O*. Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill. Crit Care Med 2002; 30: 1705-11.

12. *Kao R*, *Priestap F*, *Donner A*. To develop a regional ICU mortality prediction model during the first 24 h of ICU admission utilizing MODS and NEMS with six other independent variables from the Critical Care Information System (CCIS) Ontario. Canada. J Intensive Care 2016; DOI: https://doi.org/10.1186/s40560-016-0143-6.
13. *Altman DG*, *Vergouwe Y*, *Royston P*, *Moons KG*. Prognosis and prognostic research: validating a prognostic model. BMJ 2009; DOI: https://doi.org/10.1136/bmj.b605.
14. *Altman DG*, *Royston P*. What do we mean by validating a prognostic model? Stat Med 2000; 19: 453-73.
15. *Moons KG, Kengne AP, Grobbee DE, et al.* Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012; 98: 691-8.
16. *Bleeker SE*, *Moll HA*, *Steyerberg EW*, *et al*. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003; 56: 826-32.
17. *Cook JA*, *Collins GS*. The rise of big clinical databases. Br J Surg 2015; 102: e93-101.
18. *Riley RD*, *Ensor J*, *Snell KI*, *et al*. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016; DOI: https://doi.org/10.1136/bmj.i3140.
19. *Moons KG*, *Altman DG*, *Reitsma JB*, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015; 162: W1-73.
20. *Vergouwe Y*, *Steyerberg EW*, *Eijkemans MJ*, *Habbema JD*. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005; 58: 475-83.
21. *Hughes RA*, *Heron J*, *Sterne JA*, *Tilling K*. Accoutning for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 2019; 48: 1294-304.
22. *Debray TP*, *Vergouwe Y*, *Koffijberg H*, *Nieboer D*, *Steyerberg EW*, *Moons KG*. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015; 68: 279-89.
23. *Austin PC*, *Steyerberg EW*. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014; 33: 517-35.
24. *Steyerberg EW*, *Vergouwe Y*. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014; 35: 1925-31.
25. *Hosmer DW Jr*, *Lemeshow S*. Applied Logistic Regression. 2nd ed. NY: John Wiley & Sons; 2000 .
26. *Moreno RP, Metnitz PG, Almeida E, et al.* SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med 2005; 31: 1345-55.
27. *Higgins TL*, *Teres D*, *Copes WS*, *Nathanson BH*, *Stark M*, *Kramer AA*. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). Crit Care Med 2007; 35: 827-35.
28. *Bertolini G*, *D'Amico R*, *Nardi D*, *Tinazzi A*, *Apolone G*. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat 2000; 5: 251-3.
29. *Kramer AA*, *Zimmerman JE*. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Crit Care Med 2007; 35: 2052-6.
30. *Marcin JP*, *Romano PS*. Size matters to a model's fit. Crit Care Med 2007; 35: 2212-3.
31. *Vergouwe Y*, *Moons KG*, *Steyerberg EW*. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010; 172: 971-80.
32. *Debray TP*, *Moons KG*, *Ahmed I*, *Koffijberg H*, *Riley RD*. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med 2013; 32: 3158-80.
33. *Etthorchi-Tardy A*, *Levif M*, *Michel P*. Benchmarking: a method for continuous quality improvement in health. Healthc Policy 2012; 7: e101-19.
34. *Kiefe CI*, *Allison JJ*, *Williams OD*, *Person SD*, *Weaver MT*, *Weissman NW*. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. JAMA 2001; 285: 2871-9.
35. *Moran JL, Soloman PJ; ANZICS Centre for Outcome and Resource Evaluation (CORE) of the Australian; New Zealand Intensive Care Society (ANZICS).* Statistical process control of mortality series in the Australian and New Zealand Intensive Care Society (ANZICS) adult patient database: implications of the data generating process. BMC Med Res Methodol 2013; DOI: https://doi.org/10.1186/1471-2288-13-66 .
36. *Rasmussen TB*, *Ulrichsen SP*, *Norgaard M*. Use of risk-adjusted CUSUM charts to monitor 30-day mortality in Danish hospitals. Clin Epidemiol 2018; 10: 445-56.
37. *Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM; Emerging Risk Factors Collaboration*. Assessing risk prediction models using individual participant data from multiple studies. Am J Epidemiol 2014; 179: 621-32.
38. *Royston P*, *Parmar MK*, *Sylvester R*. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Stat Med 2004; 23: 907-26.
39. *McShane LM*, *Altman DG*, *Sauerbrei W*, *et al*. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 2005; 97: 1180-4.
40. *Moons KG*, *Altman DG*, *Vergouwe Y*, *Royston P*. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009; DOI: https://doi.org/10.1136/bmj.b606.
41. *Ferreira FL*, *Bota DP*, *Bross A*, *Melot C*, *Vincent JL*. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA 2001; 286: 1754-8.
42. *Rothen HU*, *Kung V*, *Ryser DH*, *Zurcher R*, *Regli B*. Validation of "nine equivalents of nursing manpower use score" on an independent data sample. Intensive Care Med 1999; 25: 606-11.
43. *Chen LM*, *Martin CM*, *Morrison TL*, *Sibbald WJ*. Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. Crit Care Med 1999; 27: 1999-2004.