Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.JournalofSurgicalResearch.com

Association for Academic Surgery

External Validation of Predictors of Mortality in Polytrauma Patients

Ellen R. Becker, MD,[a] *Adam D. Price, MD,*[a] *Jackson Barth, PhD,*[b,c] *Sally Hong, MIDS,*[b] *Vikas Chowdhry, MS, MBA,*[b] *Adam J. Starr, MD,*[b,d] *H. Claude Sagi, MD,*[e] *Caroline Park, MD,*[b,f] *and Michael D. Goodman, MD*[a,*]

[a] *Department of Surgery, University of Cincinnati, Cincinnati, Ohio*
[b] *TraumaCare.AI, INC, Dallas, Texas*
[c] *Department of Statistical Science, Baylor University, Waco, Texas*
[d] *Department of Orthopedic Surgery, University of Texas Southwestern Medical Center, Dallas, Texas*
[e] *Department of Orthopaedic Surgery, University of Cincinnati, Cincinnati, Ohio*
[f] *Department of Surgery, University of Texas Southwestern Medical Center, Dallas, Texas*

ARTICLE INFO

Article history:
Received 1 March 2024
Received in revised form
4 May 2024
Accepted 3 July 2024
Available online xxx

Keywords:
Machine learning
Mortality prediction
Polytrauma

ABSTRACT

Introduction: The Parkland Trauma Index of Mortality (PTIM) is an integrated, machine learning 72-h mortality prediction model that automatically extracts and analyzes demographic, laboratory, and physiological data in polytrauma patients. We hypothesized that this validated model would perform equally as well at another level 1 trauma center.
Methods: A retrospective cohort study was performed including ∼5000 adult level 1 trauma activation patients from January 2022 to September 2023. Demographics, physiologic and laboratory values were collected. First, a test set of models using PTIM clinical variables (CVs) was used as external validation, named PTIM+. Then, multiple novel mortality prediction models were developed considering all CVs designated as the Cincinnati Trauma Index of Mortality (CTIM). The statistical performance of the models was then compared.
Results: PTIM CVs were found to have similar predictive performance within the PTIM + external validation model. The highest correlating CVs used in CTIM overlapped considerably with those of the PTIM, and performance was comparable between models. Specifically, for prediction of mortality within 48 h (CTIM *versus* PTIM): positive prediction value was 35.6% *versus* 32.5%, negative prediction value was 99.6% *versus* 99.3%, sensitivity was 81.0% *versus* 82.5%, specificity was 97.3% *versus* 93.6%, and area under the curve was 0.98 *versus* 0.94.
Conclusions: This external cohort study suggests that the variables initially identified via PTIM retain their predictive ability and are accessible in a different level 1 trauma center. This work shows that a trauma center may be able to operationalize an effective predictive model without undertaking a repeated time and resource intensive process of full variable selection.

* *Corresponding author*. Department of Surgery, University of Cincinnati College of Medicine, 231 Albert Sabin Way, ML 0558, Cincinnati, OH 45267-0558.
E-mail address: goodmamd@ucmail.uc.edu (M.D. Goodman).

https://doi.org/10.1016/j.jss.2024.07.024

Introduction

Predicting mortality in polytrauma patients is imperative for prioritizing treatment plans, counseling family members, and making multidisciplinary decisions. Many scoring systems have been created in the past to aid providers with this task, including the Injury Severity Score (ISS), Acute Physiology and Chronic Health Evaluation, Trauma Score and ISS, Revised Trauma Score, and many more.[1-3] As numerous as these scores are, they come with several limitations, including having relatively low specificity,[2,4] requiring manual calculation, being static in time, relying on variables not routinely captured during trauma workflows, and being derived from stagnant models.

Incorporating machine learning to ameliorate the aforementioned limitations has grown exponentially over the past few years in all areas of medicine, and particularly in surgery. Machine learning algorithms have been implemented to aid surgeons in diagnosis,[5] treatment decisions,[6] intraoperative decision making,[7] and predicting outcomes.[8-11] One particular example of a machine learning model in this domain is the Parkland Trauma Index of Mortality (PTIM), a prediction model that automatically extracts and analyzes demographic, laboratory, and physiological data to predict mortality in trauma patients.[12]

PTIM was created to aid in surgical decision-making with the ultimate goal of improving survival within this population. It specifically predicts 72-h mortality in trauma patients starting at the12th hour of admission. It since has been internally validated and fully integrated in the real time electronic medical record (EMR) of Parkland Hospital. However, the tool has not yet been externally validated and its utility before 12 h is unexplored. In this study, it was hypothesized that the clinical variables (CVs) used for PTIM would perform similarly well at another level 1 trauma center given similar data distribution and that the PTIM CVs are an optimized set of demographic, lab and physiological data available from a standard EMR for a trauma mortality prediction model. Additionally, this study aimed to design a de novo model capable of making 72-h mortality predictions on admission, rather than 12 h postadmission.

Methods

Data

This study was completed at our academic, urban, Midwest level one trauma center. Full and partial trauma team adult activations (excluding burns) from January 2022 to September 2023 were retrospectively reviewed for demographics, physiologic and laboratory values. All CVs were extracted and then chosen based on clinical utility and availability within the EMR. Initial data extraction included ~2000 patient encounters, with a 6.6% mortality rate. Final data extraction included ~5000 patient encounters, with a 6.1% mortality rate. Patients with admission duration less than 1 h were excluded to remove the immediate deaths and patients with discharge labels as anything other than "expired," for example "hospice," were excluded. A total of 3500 CVs were considered and after clinical review and statistical validation, 49 CVs were extracted. After creating a data point for each encounter and for each hour under 12 h, observations totaled 42,154.

Machine learning

Multiple novel mortality prediction models were developed based on the retrospective data. To test our hypothesis regarding external validation of PTIM, we developed and tested a set of models using CVs identical to PTIM. This set was designated as PTIM+. For the de novo model for 1-12-h predictions, we considered the entire universe of CVs captured for full trauma team activation encounters in the EMR (~3500 unique CVs).

CVs were selected through a process involving data capture quality, clinical gestalt, and statistical methods. First, data analysts retrospectively reviewed full and partial team activations and prepared a list of CVs collected in a minimum of 10% of encounters. Second, the list was reviewed for clinical relevance and utility by a clinical team of experts who recommended a subset of those variables to be considered. Third, the refined list was utilized by data scientists for statistical validation and feature importance through machine learning methods. This set of models was designated as Cincinnati Trauma Index of Mortality (CTIM).

Both sets of models encompassed a foundational baseline employing logistic regression as well as two additional models utilizing the XGBoost algorithm. The optimal performing model type in both sets was determined to be one of the Random Forest variants. Both PTIM+ and CTIM utilize data from up to the preceding 12 h to predict 72-h mortality, and begin prediction at admission with hourly updates, producing a probability between 0 and 1.

To assess model performance, a conventional training and test set was established. Admissions beginning August 2023 and onward were considered to validate the model. The statistical performance of new models was then compared to the performance metrics previously reported for PTIM. The area under the curve (AUC) of the receiver operating characteristic curve, sensitivity, specificity, positive prediction values (PPVs) and negative prediction values (NPVs), and precision-recall curves were determined. Precision and recall are metrics used to evaluate the performance of classification models in machine learning. Precision refers to the proportion of correct items among the total items a model identifies as positive, indicating the accuracy of the positive predictions. Recall measures the proportion of actual positive items that the model correctly identifies, reflecting its ability to capture relevant items. This study was approved by the University of Cincinnati Institutional Review Board (#2022-1023) with a waiver for documentation of consent.

Results

Of the ~5000 patient encounters, the median (interquartile range) length of admission for those who survived was 40.8 (6.2, 140.0) hours compared to 8.4 (1.0, 71.5) hours for those who died, demonstrating that nearly half of all patients who die during admission, do so within the first 12 h. For patients those who survived, discharge dispositions included discharged to home without home health care (54.6%), skilled nursing facility (10.1%), rehabilitation facility (8.7%), home with home health care (7.4%), and long-term care facility (1.8%).

Of the 49 CVs extracted, 22 were utilized and generated relative importance scores based on their contribution to the model's predictive performance in CTIM, and were then compared to the respective variable importance within PTIM+ (Table 1). CVs were available in >90% of encounters apart from the following: INR (71%), arterial lactate (28%), albumin (17%), total bilirubin (16.4%), aspartate aminotransferase (16%), and base excess arterial (12%).

Performance of PTIM+ and CTIM exceeded or was comparable to the reported performance of the PTIM model on the original Parkland data. Specifically for prediction of mortality within 48 h (CTIM *versus* PTIM + *versus* PTIM): PPV was 35.6% *versus* 30.5% *versus* 32.5%, NPV was 99.6% *versus* 99.7% *versus* 99.3%, Sensitivity was 81.0% *versus* 86.1% *versus* 82.5%, Specificity was 97.3% *versus* 96.4% *versus* 93.6% and AUC was 0.98 *versus* 0.98 *versus* 0.94. Precision-recall curves for CTIM and PTIM + reported threshold cutoffs of 0.58 *versus* 0.58, respectively (Fig. 1).

To provide additional context to the performance of CTIM, an analysis was performed to determine the precision-recall of predicting mortality for lactate and base excess. Base excess alone produced precision of 81.0% and recall of 9.2%, lactate alone had precision of 81.0% and recall of 4.1%, and base excess with lactate had a precision of 81.0% and recall of 10.8% (Fig. 2).

Table 1 – CTIM *versus* PTIM + variable importance at 12 h after admission.

Clinical value	Pooling method	Variable importance (%)	
		CTIM	PTIM+
GCS	Latest	-	16.9
Best verbal response	Latest	15.5	-
Best motor response	Latest	14.4	-
Eye opening	Latest	8.8	-
Blood pressure (systolic)	Average (12 h)	7.4	1.0
Blood oxygen saturation	Minimum (12 h)	3.8	2.9
Blood pressure (systolic)	Minimum (12 h)	3.7	9.0
Arterial lactate	Latest	3.4	2.1
Mean blood oxygen saturation	Average (12 h)	2.9	10.0
Albumin	Latest	2.7	2.0
Venous lactate	Latest	2.6	2.0
Body temperature	Average (12 h)	2.3	7.9
AST	Latest	2.1	1.4
INR	Latest	1.9	1.6
Venous base excess	Latest	1.9	4.4
Blood pressure (systolic)	Latest	1.9	3.2
Body temperature	Minimum (12 h)	1.9	10.0
Age at admission	N/A	1.7	2.2
Total bilirubin	Latest	1.7	2.2
Body temperature	Maximum (12 h)	1.7	2.0
Heart rate	Maximum (12 h)	1.7	1.4
Blood pressure (systolic)	Maximum (12 h)	1.7	2.1
Body temperature	Latest	1.6	2.4
Platelet count	Latest	1.6	3.0
Heart rate	Latest	1.5	2.4
Potassium	Latest	1.5	2.2
Arterial base excess	Latest	1.4	2.0
Respirations	Maximum (12 h)	1.4	-
Blood oxygen saturation	Maximum (12 h)	1.2	1.9
Blood oxygen saturation	Latest	1.1	1.5
Heart rate	Average (12 h)	1.1	-
Respirations	Average (12 h)	1.1	-
Hemoglobin concentration	Latest	0.9	-

Blank value indicated CV was not in model and/or not found in feature importance. Variable importance is calculated by dividing individual score by total score.
AST = aspartate aminotransferase.

Discussion

PTIM+ and CTIM use machine learning to predict mortality in polytrauma patients using readily available CVs commonly captured in daily trauma workflows within the EMR. This means that these models can be integrated with the EMR in real time without requiring any additional manual provider input of data – a step which is usually an insurmountable barrier to clinical adoption of AI tools. Since the performance of PTIM+ and CTIM exceeded or was comparable to the reported performance of the PTIM model on the original Parkland data, and there is a significant overlap of CV features between PTIM and CTIM with similar if not improved predictive performance, this affirmed the hypothesis of our study and provided external validation for the PTIM model. The improved model performance is credited to a larger training data set with more granular features, an improved machine learning algorithm using XGBoost, and a more strategic, streamlined approach for feature engineering. Finally, with these more advanced models, PTIM+ and CTIM successfully created a de novo prediction model that enabled the algorithm to begin making 72-h mortality predictions for patients on admission, rather than 12 h post admission.

Prior to machine learning, various scoring systems as previously mentioned were used to predict mortality in trauma patients. Studies have estimated an ISS sensitivity and specificity of 79%-81% and 61%-74%, and an Revised Trauma Score sensitivity and specificity of 81%-87% and 60%-80%, respectively.[2,4] In contrast, PTIM first began with data collection from 2009 to 2014, and was validated on encounters in 2015-2016.[12] The original prediction model also utilized 23 variables with varying importance and performed with high sensitivity and specificity. Specifically, the model was only incorrect 11 out of 1608 time

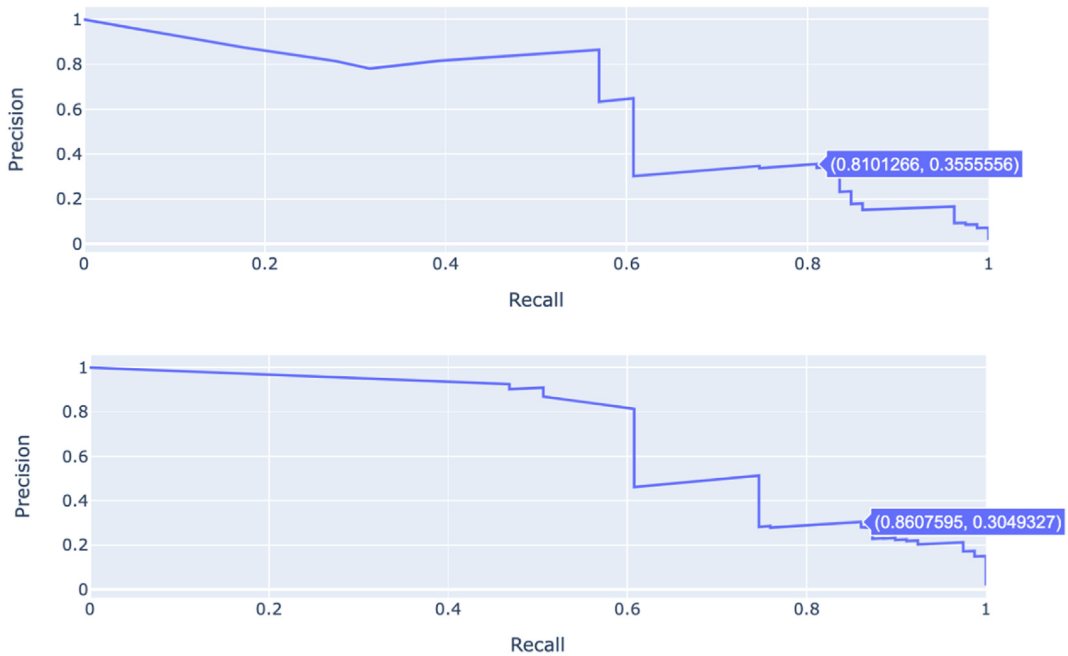

Fig. 1 – Precision-recall curves for CTIM (top) and PTIM+ (bottom). Coordinates indicate CTIM to have a precision of 81.0% and a recall of 35.6%, and PTIM + to have a precision of 86.1% and a recall of 30.5%.

intervals during preclinical use. Clinical use of PTIM demonstrated a sensitivity of 86.9%, specificity of 94.7%, PPV 33.3%, NPV 99.6% and AUC of 0.97.[12] The current study similarly found these CVs, with the expansion of GCS, to again create the most accurate prediction model.

Updates to PTIM were published using a retrospective cohort from 2020 to 2022, which included a direct comparison of PTIM to other scoring systems and variables used to predict mortality, including ISS. It was found that PTIM and ISS had only a weak positive correlation ($r = 0.315$, $P < 0.001$), and importantly, that operative mortality was increased in the high-risk PTIM cohort (9.1% *versus* 1.2%, $P < 0.001$), but no worse in patients with an ISS greater than 15 (4.2% *versus* 1.4%, $P = 0.09$).[13] Our study also aimed to provide additional context to the performance of the CTIM model through comparison using lactate and base excess, being known measures of clinical and physiological instability.[14-16] These models showed similar precision to CTIM, but with recall rates roughly one-third those of CTIM, being representative of the ability of the model to detect a true positive. Ideally, in future work, more rigorous methods to measure provider ability to predict mortality and compare it with capabilities of a machine learning derived algorithm could be utilized.

Future directions for CTIM will include a prospective, real-time evaluation of the clinical and operational effectiveness of these models by embedding them in trauma and critical care electronic health record workflows. For PTIM, this has already been accomplished and from a prior report, it was found that 35 of 40 providers used PTIM in decision-making, 77% for treatment planning and 66% for surgical intervention timing.[17] Furthermore, 61% found it was easy to use and 56% agreed it improved efficiency in assessing potential mortality.[17] Simultaneously as CTIM is incorporated into provider workflow at the University of Cincinnati, additional sites will also begin to undertake full variable selection and validation, further solidifying and improving the 23-variable algorithm.

This study has the following limitations. First, data collected was from two centers thus far, both being urban level one trauma centers. Second, the algorithm is limited by variables available in the EMR, which inherently have entry errors and may not capture the granularity of moment-to-moment changes in variables such as vital signs. Last, the

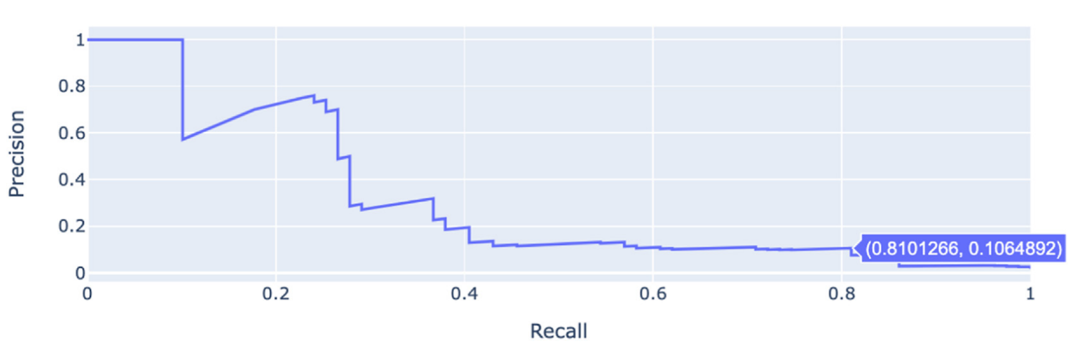

Fig. 2 – Base-excess and lactate precision-recall curve. Coordinates denote precision to be 81.0% and recall to be 10.6%.

PTIM, PTIM+, and CTIM data were all analyzed and validated by the creators of the machine learning algorithms.

In conclusion, this external cohort study suggests that the model features initially identified via PTIM retain their predictive ability for mortality in a different level 1 trauma center. Additionally, the variables included in the original PTIM dataset are accessible and analyzable in another level 1 trauma center, demonstrating generalizability of these predictors. Finally, the prediction model can be expanded to predict 72-h mortality as early as 1 h after admission. This work shows promise that a trauma center may be able operationalize an effective predictive model without undertaking a repeated time and resource intensive process of full variable selection and can be reproduced at additional trauma centers.

Study Type

Retrospective data analysis.

Level of Evidence

III.

Disclosure

Dr Park and Dr Starr are clinical advisors to TraumaCare.AI, INC. Barth, Hong, Park, Starr and Chowdhry have received equity from TraumaCare.AI.

Funding

None.

Meeting Presentation

19th Annual Academic Surgical Congress Quickshot.

CRediT authorship contribution statement

Ellen R. Becker: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Adam D. Price:** Writing – review & editing, Project administration. **Jackson Barth:** Writing – review & editing, Formal analysis, Conceptualization. **Sally Hong:** Validation, Methodology, Formal analysis. **Vikas Chowdhry:** Validation, Methodology, Formal analysis, Conceptualization. **Adam J. Starr:** Writing – review & editing, Methodology, Conceptualization. **H. Claude Sagi:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Caroline Park:** Writing – review & editing, Methodology, Conceptualization. **Michael D. Goodman:** Writing – review & editing, Supervision, Methodology, Conceptualization.

REFERENCES

1. Mijaljica DR, Gregoric P, Ivancevic N, Pavlovic V, Jovanovic B, Djukic V. Predicting mortality in severe polytrauma with limited resources. *Ulus Travma Acil Cerrahi Derg*. 2022;28:1404–1411.
2. Milton M, Engelbrecht A, Geyser M. Predicting mortality in trauma patients - a retrospective comparison of the performance of six scoring systems applied to polytrauma patients from the emergency centre of a South African central hospital. *Afr J Emerg Med*. 2021;11:453–458.
3. Pape HC, Champion HR. Patient assessment in polytrauma: current trends rely on multiple parameters to improve the prediction of complications and mortality. *Injury*. 2015;46:1875–1877.
4. Merchant AAH, Shaukat N, Ashraf N, et al. Which curve is better? A comparative analysis of trauma scoring systems in a South Asian country. *Trauma Surg Acute Care Open*. 2023;8:e001171.
5. Maki S, Furuya T, Inoue M, et al. Machine learning and deep learning in spinal injury: a narrative review of algorithms in diagnosis and prognosis. *J Clin Med*. 2024;13:705.
6. Dandurand C, Fallah N, Öner CF, et al. Predictive algorithm for surgery recommendation in thoracolumbar burst fractures without neurological deficits. *Glob Spine J*. 2024;14:56S–61S.
7. Andriollo L, Picchi A, Sangaletti R, et al. The role of artificial intelligence in anterior cruciate ligament injuries: current concepts and future perspectives. *Healthcare (Basel)*. 2024;12:300.
8. Wang SY, Barrette LX, Ng JJ, et al. Predicting reoperation and readmission for head and neck free flap patients using machine learning. *Head Neck*. 2024;46:1999–2009.
9. Yin AA, Zhang X, He YL, et al. Machine learning prediction models for in-hospital postoperative functional outcome after moderate-to-severe traumatic brain injury. *Eur J Trauma Emerg Surg*. 2024.
10. Li B, Shaikh F, Zamzam A, Syed MH, Abdin R, Qadura M. A machine learning algorithm for peripheral artery disease prognosis using biomarker data. *iScience*. 2024;27:109081.
11. Zhang W, Wang S. Machine learning developed an intratumor heterogeneity signature for predicting prognosis and immunotherapy benefits in skin cutaneous melanoma. *Melanoma Res*. 2024;34:215–224.
12. Starr AJ, Julka M, Nethi A, et al. Parkland trauma Index of mortality: real-time predictive model for trauma patients. *J Orthop Trauma*. 2022;36:280–286.
13. Tiziani S, Hinkle AJ, Mesarick EC, et al. Parkland trauma Index of mortality in orthopaedic trauma patients: an initial report. *J Orthop Trauma*. 2023;37:S23–S27.
14. Kosaki Y, Hongo T, Hayakawa M, et al. Association of initial lactate levels and red blood cell transfusion strategy with outcomes after severe trauma: a post hoc analysis of the RESTRIC trial. *World J Emerg Surg*. 2024;19:1.
15. Lee GR, Kim HS, Chung YJ, Kim EY. Venous-to-arterial carbon dioxide tension difference measurement as a useful predictor of patient prognosis after major surgery. *Asian J Surg*. 2024;47:2152–2160.
16. Mayor J, Gräff P, Birgel V, et al. Do admission glucose levels independently predict coagulopathy in multiple trauma patients? A retrospective cohort analysis. *Eur J Trauma Emerg Surg*. 2024.
17. Park C, Loza-Avalos SE, Harvey J, et al. A real-time automated machine learning algorithm for predicting mortality in trauma patients: survey says it's ready for prime-time. *Am Surg*. 2023;90:655–661.