Has Machine Learning Made a Difference Yet? A Retrospective Analysis of In-Hospital Mortality Prediction

Michael John Patton & Matthew Might

Hugh Kaul Precision Medicine Institute, University of Alabama at Birmingham, AL, USA

1 Introduction

Since the 1980s, medical experts have designed and refined dozens of point-based scoring systems for predicting critical care outcomes ranging from mortality and length of hospital stay to sepsis and septic shock.[1–3] With significant advances in computational infrastructure, machine learning (ML) methods, and wide-spread adoption of electronic health records (EHR), outcome prediction research underwent rapid expansion from 2010-2020. Despite an exponential increase in research volume and diversity of methods, the overall range of reported model performance for in-hospital mortality prediction has surprisingly widened. Moreover, peak model performance from research conducted within the past 5 years has shown only marginal improvement over research conducted over 20 years ago. To assess the progress of these new methods, we conducted a retrospective analysis from 1980-2022 of published severity score and ML models for in-hospital mortality prediction.

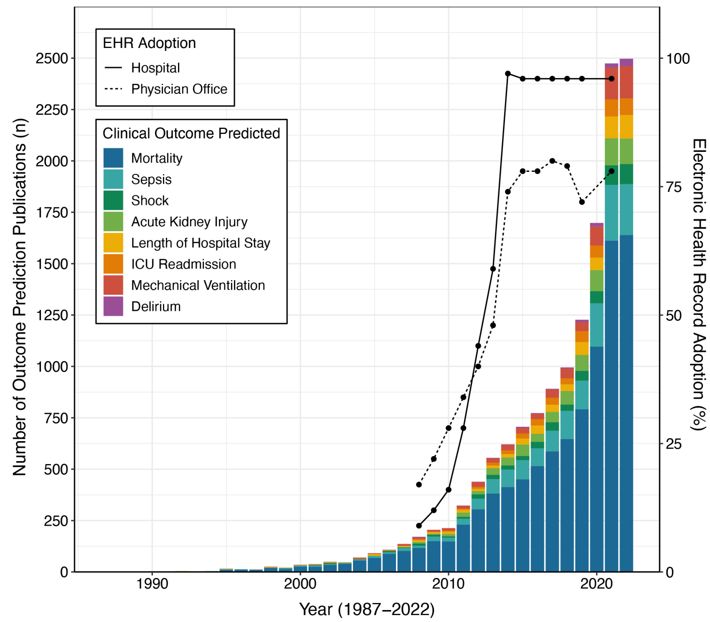

Figure 1: Time-Series of US electronic health record adoption and concurrent clinical outcome prediction publications from 1980-2022. Electronic health record (EHR) adoption data visualized from reference cited.(Abbreviations: United States of America, US)[4]

2 Methods

We performed a systematic query of MEDLINE for predictive outcome research from 1980-2022. Published research was manually curated and selected based on 3 search criterion: 1) in-hospital mortality outcome models, 2) training data using adult inpatients cohorts, 3) model performance evaluation using an area under the receiver operating characteristic curve (AUROC, formerly the c-statistic). Studies were then categorized by severity score, and/or ML algorithm.

3 Results

From 2008 to 2014, adoption of electronic health records (EHR) in American hospitals grew exponentially from 9% to 96%. During this expansion, the number of outcome prediction publications from 2008-2011 alone surpassed the sum total of the previous two-decades.

From the 14,465 manuscripts published from 1980-2022, a total of 66% (n=9,602) were identified as mortality prediction studies. After filtering for in-hospital mortality as the predicted outcome and AUROC as the performance metric(s) reported, 77 publications were selected for further evaluation.

3.1 Severity Score Performance

The top performing severity score-based models from 1980-2000, 2000-2010, and 2010-2022 were

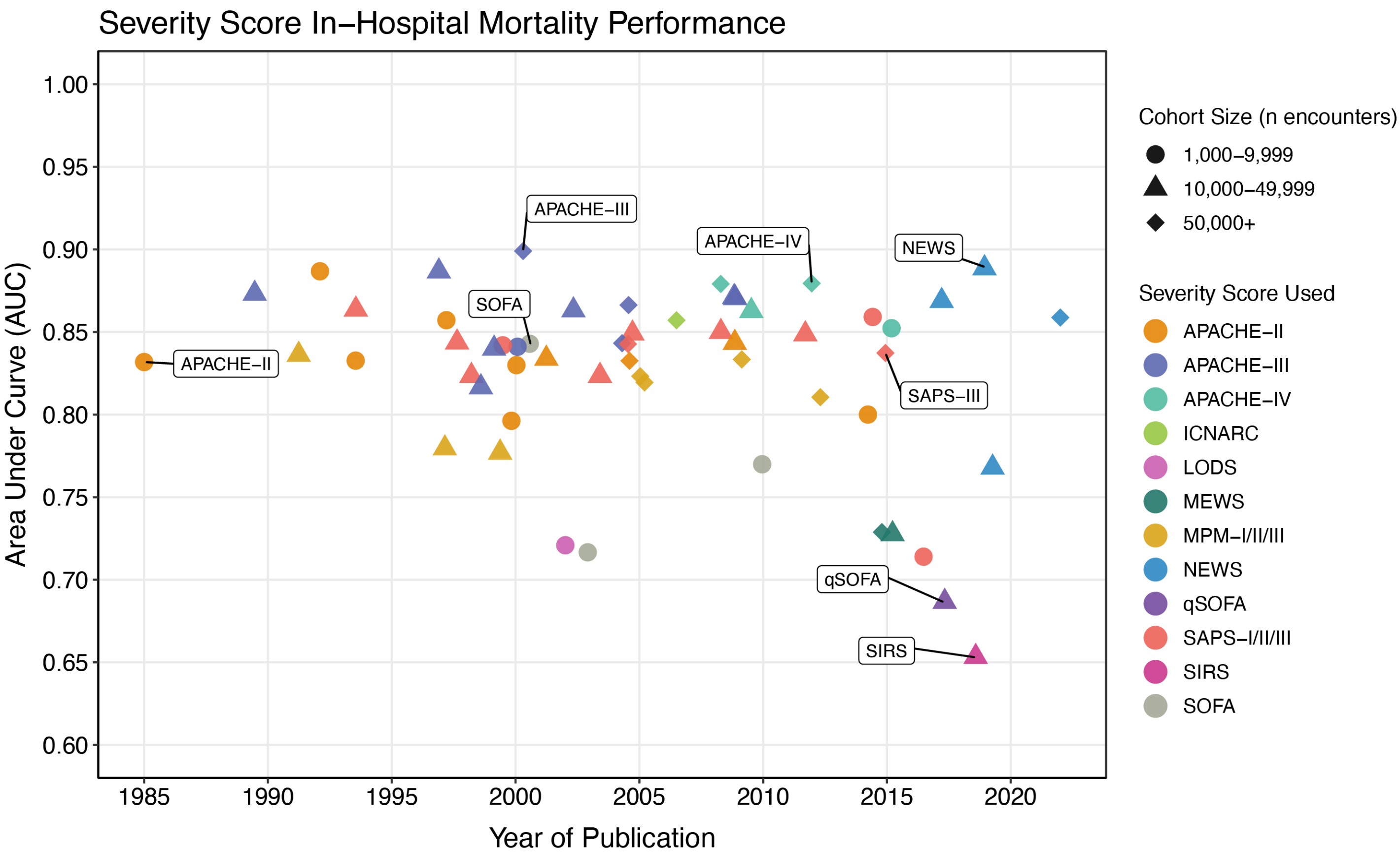

Figure 2: In-Hospital Mortality Prediction with Severity Score Models. Severity score model performance for predicting in-hospital mortality for adult inpatient encounters from 1985-2022. Abbreviations: Sequential Organ Failure Score (SOFA), quick SOFA (qSOFA), Systemic Inflammatory Response Syndrome (SIRS), Logistic Organ Dysfunction Score (LODS), Simplified Acute Physiology Score (SAPS), Modified Early Warning Score (MEWS), National Early Warning Score (NEWS), Mortality Prediction Model (MPM), Intensive Care National Audit and Research Center (ICNARC), Acute physiological assessment and chronic health evaluation (APACHE). Data visualized from references cited.[1,2,5–37]

the Acute Physiology and Chronic Health Evaluation III (APACHE-III) score (AUROC range: 0.82-0.90), the APACHE-IV score (AUROC: 0.88), and New Early Warning (NEWS) score (AUROC range: 0.77-0.89), respectively for studies including >1000 patient records (Figure 2). Peak performance for in-hospital mortality prediction using severity scores was observed in 1999 (APACHE-III AUROC: 0.90; Figure 2). We observed a marginal decrease in peak performance (-0.02 AUROC) from the APACHE-III scores (n=18 variables) to the APACHE-IV scores (n=129 variables) despite an increase of 111 scoring variables (Figure 2).[15,33]

3.2 Machine Learning Performance

ML model performance for in-hospital mortality prediction as binary outcome exhibited a wide range in performance (AUROC range=0.57-0.95). Neural network and random forest methods consistently exhibited peak performance (AUROC: ≥0.90); however, traditional regression methods that utilized novel data-types such as medical free-text and regession coefficients from time-series laboratory measurement trends as model variables had comparable performance (Figure 3, Marafino et al, AUROC: 0.92). Significant differences in model performance on derivation (AUROC: 0.95) and validation data (AUROC: 0.81) were observed (Figure 3, Meyer et al.).

4 Conclusion

This 40-year retrospective analysis of in-hospital mortality prediction research revealed a widening gap of reported ML performance but significant improvement with utilization of medical free text and time-series data as model variables.

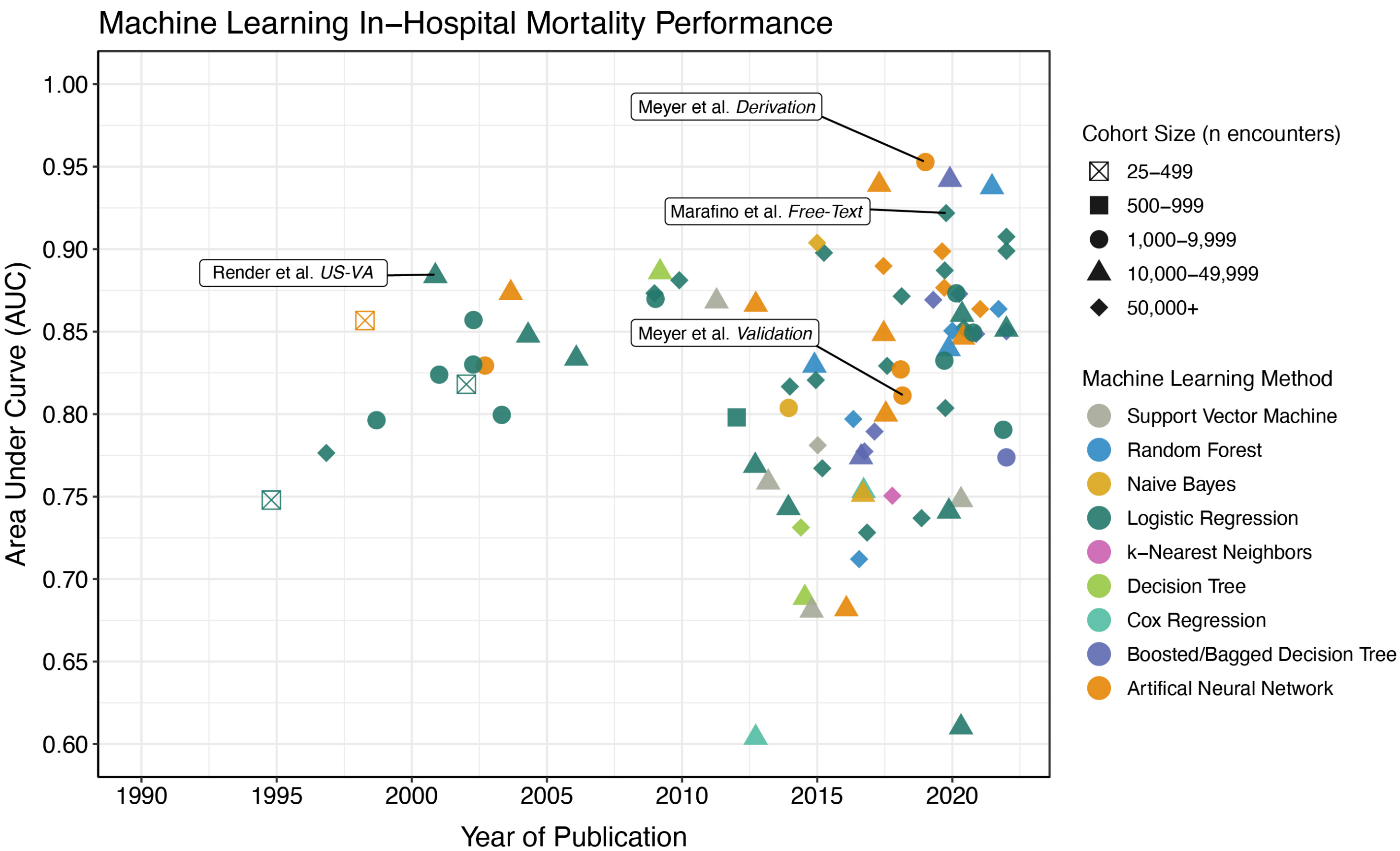

Figure 3: In-Hospital Mortality Prediction with Machine Learning Models. Machine learning model performance for predicting in-hospital mortality for adult inpatient encounters (ICU and general admission) from 1990-2022. Specific studies are high-lighted, with "Derivation" and "Validation" labels referring to study specific training/testing data and out-of-sample validation data, respectively. "Free-Text" labels refer to newer machine learning studies that have incorporated of these respective data-types as model features. Abbreviations: United States Veterans Association data (US-VA). Data visualized was collected from references cited.[22,29,31,36,38–83]

References

1. Knaus WA, Draper EA, Wagner DP, and Zimmerman JE. APACHE II: a severity of disease classification system. Critical Care Medicine 1985;13:818–29.
2. Le Gall JR, Lemeshow S, and Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. The Journal of the American Medical Association 1993;270:2957–63.
3. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Medicine 1996;22:707–10.
4. National Trends in Hospital and Physician Adoption of Electronic Health Records from the American Hospital Association (AHA) Annual Survey Information Technology Supplement, 2008-present |HealthIT.gov. WEBSITE.
5. Harrison DA, Brady AR, Parry GJ, Carpenter JR, and Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Critical Care Medicine 2006;34:1378–88.
6. Moreno R, Vincent JL, Matos R, et al. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. Working Group on Sepsis related Problems of the ESICM. Intensive Care Medicine 1999;25:686–96.
7. Brinkman S, Bakhshi-Raiez F, Abu-Hanna A, et al. External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute

Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. Journal of Critical Care 2011;26:105.e11–8.

8. Harrison DA and Rowan KM. Outcome prediction in critical care: the ICNARC model. Current Opinion in Critical Care 2008;14:506–12.

9. Redfern OC, Pimentel MAF, Prytherch D, et al. Predicting in-hospital mortality and unanticipated admissions to the intensive care unit using routinely collected blood tests and vital signs: Development and validation of a multivariable model. Resuscitation 2018;133:75–81.

10. Rivera-Fernández R, Vázquez-Mata G, Bravo M, et al. The Apache III prognostic system: customized mortality predictions for Spanish ICU patients. Intensive Care Medicine 1998;24:574–81.

11. Beck DH, Smith GB, Pappachan JV, and Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Medicine 2003;29:249–56.

12. Poole D, Rossi C, Anghileri A, et al. External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units. Intensive Care Medicine 2009;35:1916–24.

13. Kramer AA, Higgins TL, and Zimmerman JE. Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: implications for national benchmarking*. Critical Care Medicine 2014;42:544–53.

14. Oh TE, Hutchinson R, Short S, Buckley T, Lin E, and Leung D. Verification of the Acute Physiology and Chronic Health Evaluation scoring system in a Hong Kong intensive care unit. Critical Care Medicine 1993;21:698–705.

15. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619–36.

16. Higgins TL, Kramer AA, Nathanson BH, Copes W, Stark M, and Teres D. Prospective validation of the intensive care unit admission Mortality Probability Model (MPM0-III). Critical Care Medicine 2009;37:1619–23.

17. Rowan KM, Kerr JH, Major E, McPherson K, Short A, and Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland–II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. BMJ (Clinical Research Ed.) 1993;307:977–81.

18. Moreno R, Miranda DR, Fidler V, and Van Schilfgaarde R. Evaluation of two outcome prediction models on an independent database. Critical Care Medicine 1998;26:50–61.

19. Sirio CA, Shepardson LB, Rotondi AJ, et al. Community-wide assessment of intensive care outcomes using a physiologically based prognostic measure: implications for critical care delivery from Cleveland Health Quality Choice. Chest 1999;115:793–801.

20. Engerström L, Nolin T, M C, et al. Impact of Missing Physiologic Data on Performance of the Simplified Acute Physiology Score 3 Risk-Prediction Model. Critical Care Medicine 2017;45:2006–13.

21. Lee H, Shon YJ, Kim H, Paik H, and Park HP. Validation of the APACHE IV model and its comparison with the APACHE II, SAPS 3, and Korean SAPS 3 models for the prediction of hospital mortality in a Korean surgical intensive care unit. Korean Journal of Anesthesiology 2014;67:115–22.

22. Duke GJ, Santamaria J, Shann F, et al. Critical care outcome prediction equation (COPE) for adult intensive care. Critical Care and Resuscitation 2008;10:41.

23. Timsit JF, Fosse JP, Troché G, et al. Calibration and discrimination by daily Logistic Organ Dysfunction scoring comparatively with daily Sequential Organ Failure Assessment scoring for predicting hospital mortality in critically ill patients. Critical Care Medicine 2002;30:2003–13.

24. Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, and Knaus WA. Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. Critical Care Medicine 1998;26:1317–26.

25. Moreno RP, Metnitz PGH, 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 Medicine 2005;31:1345–55.

26. Wong DT, Crofts SL, Gomez M, McGuire GP, and Byrick RJ. Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients. Critical Care Medicine 1995;23:1177–83.

27. Markgraf R, Deutschinoff G, Pientka L, and Scholten T. Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. Critical Care Medicine 2000;28:26–33.

28. Livingston BM, MacKirdy FN, Howie JC, Jones R, and Norrie JD. Assessment of the performance of five intensive care scoring models within a large Scottish database. Critical Care Medicine 2000;28:1820–7.

29. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. The Lancet. Respiratory medicine 2018;6:905–14.

30. Liu VX, Lu Y, Carey KA, et al. Comparison of Early Warning Scoring Systems for Hospitalized Patients With and Without Infection at Risk for In-Hospital Mortality and Transfer to the Intensive Care Unit. JAMA network open 2020;3:e205191.

31. Kim S, Kim W, and Park RW. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. Healthcare informatics research 2011;17:232–43.

32. Cárdenas-Turanzas M, Ensor J, Wakefield C, et al. Cross-validation of a Sequential Organ Failure Assessment score-based model to predict mortality in patients with cancer admitted to the intensive care unit. Journal of Critical Care 2012;27:673–80.

33. Zimmerman JE, Kramer AA, McNair DS, and Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Critical Care Medicine 2006;34:1297–310.

34. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, and Kramer AA. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). Critical Care Medicine 2007;35:827–35.

35. Churpek MM, Snyder A, Han X, et al. Quick Sepsis-related Organ Failure Assessment, Systemic Inflammatory Response Syndrome, and Early Warning Scores for Detecting Clinical Deterioration in Infected Patients outside the Intensive Care Unit. American Journal of Respiratory and Critical Care Medicine 2017;195:906–11.

36. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, and Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical Care Medicine 2016;44:368–74.

37. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, and Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. The Journal of the American Medical Association 1993;270:2478–86.

38. Adrie C, Francais A, Alvarez-Gonzalez A, et al. Model for predicting short-term mortality of severe sepsis. Critical Care 2009;13:R72.

39. Render ML, Welsh DE, Kollef M, et al. Automated computerized intensive care unit severity of illness measure in the Department of Veterans Affairs: preliminary results. SISVistA Investigators. Scrutiny of ICU Severity Veterans Health Sysyems Technology Architecture. Critical Care Medicine 2000;28:3540–6.

40. Moran JL, Solomon PJ, Outcome AC for, Australian RE(of, and (ANZICS) NZICS. Fixed effects modelling for provider mortality outcomes: Analysis of the Australia and New Zealand Intensive Care Society (ANZICS) Adult Patient Data-base. Plos One 2014;9:e102297.

41. Dybowski R, Weller P, Chang R, and Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. The Lancet 1996;347:1146–50.

42. Wang Y, Chen W, Heard K, et al. Mortality Prediction in ICUs Using A Novel Time-Slicing Cox Regression Method. AMIA Annual Symposium Proceedings 2015;2015:1289–95.

43. Hunziker S, Celi LA, Lee J, and Howell MD. Red cell distribution width improves the simplified acute physiology score for risk prediction in unselected critically ill patients. Critical Care 2012;16:R89.

44. Moran JL, Bristow P, Solomon PJ, et al. Mortality and length-of-stay outcomes, 1993-2003, in the binational Australian and New Zealand intensive care adult patient database. Critical Care Medicine 2008;36:46–61.

45. Angus DC, Linde-Zwirble WT, Sirio CA, et al. The effect of managed care on ICU length of stay: implications for medicare. The Journal of the American Medical Association 1996;276:1075–82.

46. Clermont G, Angus DC, DiRusso SM, Griffin M, and Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Critical Care Medicine 2001;29:291–6.

47. Render ML, Kim HM, Welsh DE, et al. Automated intensive care unit risk adjustment: results from a National Veterans Affairs study. Critical Care Medicine 2003;31:1638–46.

48. Minne L, Toma T, Jonge E de, and Abu-Hanna A. Assessing and combining repeated prognosis of physicians and temporal models in the intensive care. Artificial Intelligence in Medicine 2013;57:111–7.

49. Churpek MM, Yuen TC, and Edelson DP. Predicting clinical deterioration in the hospital: the impact of outcome selection. Resuscitation 2013;84:564–8.

50. Keizer NF de, Bonsel GJ, Goldfad C, and Rowan KM. The added value that increasing levels of diagnostic information provide in prognostic models to estimate hospital mortality for adult intensive care patients. Intensive Care Medicine 2000;26:577–84.

51. Duke GJ, Barker A, Rasekaba T, Hutchinson A, and Santamaria JD. Development and validation of the critical care outcome prediction equation, version 4. Critical Care and Resuscitation 2013;15:191–7.

52. Render ML, Deddens J, Freyberg R, et al. Veterans Affairs intensive care unit risk adjustment model: validation, updating, recalibration. Critical Care Medicine 2008;36:1031–42.

53. Timsit JF, Fosse JP, Troché G, et al. Accuracy of a composite score using daily SAPS II and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 h. Intensive Care Medicine 2001;27:1012–21.

54. Markgraf R, Deutschinoff G, Pientka L, Scholten T, and Lorenz C. Performance of the score systems Acute Physiology and Chronic Health Evaluation II and III at an interdisciplinary intensive care unit, after customization. Critical Care 2001;5:31–6.

55. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). The Journal of the American Medical Association 2016;315:762–74.

56. Raith EP, Udy AA, Bailey M, et al. Prognostic Accuracy of the SOFA Score, SIRS Criteria, and qSOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit. The Journal of the American Medical Association 2017;317:290–300.

57. Awad A, Bader-El-Den M, McNicholas J, and Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensem-

ble learning approach. International Journal of Medical Informatics 2017;108:185–95.

58. Silva A, Cortez P, Santos MF, Gomes L, and Neves J. Mortality assessment in intensive care units via adverse events using artificial neural networks. Artificial Intelligence in Medicine 2006;36:223–34.

59. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, and Rashidi P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Scientific Reports 2019;9:1879.

60. Kim SY, Kim S, Cho J, et al. A deep learning model for real-time mortality prediction in critically ill children. Critical Care 2019;23:279.

61. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, and Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Critical Care 2019;23:64.

62. Brajer N, Cozzi B, Gao M, et al. Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. JAMA network open 2020;3:e1920733.

63. Marafino BJ, Park M, Davies JM, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. JAMA network open 2018;1:e185097.

64. Holmgren G, Andersson P, Jakobsson A, and Frigyesi A. Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions. Journal of intensive care 2019;7:44.

65. Parreco JP, Hidalgo AE, Badilla AD, Ilyas O, and Rattan R. Predicting central line-associated bloodstream infections and mortality using supervised machine learning. Journal of Critical Care 2018;45:156–62.

66. Wong RSY and Ismail NA. An application of bayesian approach in modeling risk of death in an intensive care unit. Plos One 2016;11:e0151949.

67. He F, Page JH, Weinberg KR, and Mishra A. The Development and Validation of Simplified Machine Learning Algorithms to Predict Prognosis of Hospitalized Patients With COVID-19: Multicenter, Retrospective Study. Journal of Medical Internet Research 2022;24:e31549.

68. Tang F, Xiao C, Wang F, and Zhou J. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA open 2018;1:87–98.

69. Hug CW and Szolovits P. ICU acuity: real-time models versus daily models. AMIA Annual Symposium Proceedings 2009;2009:260–4.

70. Iwase S, Nakada TA, Shimada T, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. Scientific Reports 2022;12:12912.

71. Ren Y, Loftus TJ, Datta S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. JAMA network open 2022;5:e2211973.

72. Hsu YT, He YT, Ting CK, Tsou MY, Tang GJ, and Pu C. Administrative and claims data help predict patient mortality in intensive care units by logistic regression: A nationwide database study. BioMed research international 2020;2020:9076739.

73. Elhazmi A, Al-Omari A, Sallam H, et al. Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU. Journal of infection and public health 2022;15:826–34.

74. Higgins TL, Freeseman-Freeman L, Stark MM, and Henson KN. Benchmarking inpatient mortality using electronic medical record data: A retrospective, multicenter analytical observational study. Critical Care Medicine 2022;50:543–53.

75. Keller MB, Wang J, Nason M, Warner S, Follmann D, and Kadri SS. Preintubation Sequential Organ Failure Assessment Score for Predicting COVID-19 Mortality: External Validation Using Electronic Health Record From 86 U.S. Healthcare Systems to Appraise Current

Ventilator Triage Algorithms. Critical Care Medicine 2022;50:1051–62.

76. Raffa JD, Johnson AEW, O'Brien Z, et al. The global open source severity of illness score (GO). Critical Care Medicine 2022;50:1040–50.

77. Zhang Z and Hong Y. Development of a novel score for the prediction of hospital mortality in patients with severe sepsis: the use of electronic healthcare records with LA regression. Oncotarget 2017;8:49637–45.

78. Xu Y, Trivedi A, Becker N, et al. Machine learning-based derivation and external validation of a tool to predict death and development of organ failure in hospitalized patients with COVID-19. Scientific Reports 2022;12:16913.

79. Reina Reina A, Barrera JM, Valdivieso B, Gas ME, Maté A, and Trujillo JC. Machine learning model from a Spanish cohort for prediction of SARS-COV-2 mortality risk and critical patients. Scientific Reports 2022;12:5723.

80. Daly K, Beale R, and Chang RW. Reduction in mortality after inappropriate early discharge from intensive care unit: logistic regression triage model. BMJ (Clinical Research Ed.) 2001;322:1274–6.

81. Edelson M and Kuo TT. Generalizable prediction of COVID-19 mortality on worldwide patient data. JAMIA open 2022;5:ooac036.

82. Houthooft R, Ruyssinck J, Herten J van der, et al. Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artificial Intelligence in Medicine 2015;63:191–207.

83. Sigakis MJG, Bittner EA, and Wanderer JP. Validation of a risk stratification index and risk quantification index for predicting patient outcomes: in-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay. Anesthesiology 2013;119:525–40.