

No-Code Training of Machine Learning Models using Intensive Care Unit Serology Hospital Data to Predict Survival outcomes of Hospital Surgery

John Yee Nok Ng¹, David Jijeong Lee¹, Sneha Nayak Chhetri¹

¹ School of Medicine and Dentistry, Griffith University, Gold Coast, Queensland, Australia

How to Cite this Article (Vancouver Style)

Ng JYN, Lee DJ, Chhetri SN. No-Code Training of Machine Learning Models using Intensive Care Unit Serology Hospital Data to Predict Survival outcomes of Hospital Surgery. Med J Ins Sc Tech. 2026;1(1):10–20. DOI: <https://doi.org/10.68003/mjiosat.v1i16>.

Abstract

Introduction: Surgery and prolonged admission in intensive care units (ICU) can introduce complications such as sepsis, iatrogenic infections and other adverse events. Early recognition of serology laboratory markers by AI-assisted tools could identify high-risk patients to guide proactive decision-making. This study aimed to develop and evaluate no-code AI prediction models to estimate postoperative survival in ICU-admitted patients.

Materials and Methods: Open-source, de-identified ICU data from The Zigong Fourth People's Hospital, China, available on Physio.net2 were analysed using risk-ratios to identify a basket of high-risk serological markers. These markers were then used as features to develop self-learning decision-tree models to predict patient survival outcomes based on their serological data.

Results: Multiple laboratory markers were associated with mortality outcomes. Our initial Decision-Tree model had a prediction accuracy of 56%, with similar results across Neural Networks (58.0%) and Random Forest models (57.9%). Performance across all models was consistent (~58.0%) against the Boston-Symile-MIMIC-database.

Conclusion: This assessment of AI learning models demonstrated only modest accuracy (56-58%) in clinical prediction, limiting its clinical applicability. This study, however, did not leverage larger cloud architectures like Amazon Web Services or Google Cloud which could allow significantly larger feature selection. Predictive performance could be improved in future studies by scaling to large-scale cloud computing resources so that feature sets are not limited by on-site infrastructure constraints.

Keywords: Postoperative Complications, Intensive Care Units, Artificial Intelligence, Biomarkers, Sepsis, Survival, Outcomes, Machine Learning.

Article identifiers

eISSN: 3149-7497

Article ID: MJIOSAT-v1i1.a01

Pages: 10-20

DOI: 10.68003/mjiosat.v1i16



Scan to Access

Article History

Received: October 15, 2025

Revised: January 2, 2026

Acceptance: January 30, 2026

Published: April 30, 2026

Corresponding Author

Sneha Nayak Chhetri

School of Medicine and Dentistry, Griffith University – Gold Coast Campus, Southport QLD 4215, Australia

Email: sneha.nayakchhetri@griffithuni.edu.au

ORCID ID: [0009-0004-1281-4412](https://orcid.org/0009-0004-1281-4412)

Introduction

Postoperative infection and delayed tissue healing continue to pose substantial challenges in postoperative care, contributing significantly to morbidity, prolonged intensive care unit (ICU) admission, and mortality.¹⁻⁴ Postoperative complications may progress to severe systemic manifestations-including systemic inflammatory response syndrome (SIRS), sepsis, and multiorgan dysfunction-that worsen patient outcomes and impose a considerable burden on health care systems.⁵⁻⁸ Early recognition of the

physiological factors and laboratory markers that precede such severe complications is therefore critical to improving prognosis.^{6,9}

Although various risk assessment frameworks are in use, their utility is often constrained by the requirement for expert clinical interpretation. In practice, the increasing volume and complexity of patient data place considerable cognitive demands on physicians, resulting in variability

in decision making and limiting the reliability of these traditional tools.¹⁰

Advances in artificial intelligence (AI) may offer unprecedented capacity to interrogate large-scale datasets and generate more accurate, dynamic, and personalized risk stratification models. With the rapid evolution of both AI methodologies and the availability of open-source data, the development of AI-assisted predictive tools could assist many physicians in the clinic. AI diagnostic tools may also fundamentally enhance decision making by enabling earlier identification of high-risk patients and guiding targeted, proactive management strategies.¹²

This study aimed to develop and assess a preoperative, AI-assisted clinical prediction model to estimate postoperative survival in ICU-admitted surgical patients. Such a model could be a useful clinical tool to assess whether surgical intervention leads to a higher risk of death (poorer outcomes) or survival (better outcomes) in high-risk patients.

We hypothesized that an AI training model could accurately and definitively predict survival or death outcomes for a patient who underwent surgical intervention (based on the patient's health records). Under the null hypothesis, our AI model's predictive performance is no better than chance, indicating that the observed results are consistent with a random distribution.

Materials and Methods

Public hospital data were sourced from the open-source website Research Resource for Complex Physiologic Signals (physio.net).²⁻⁴ Amongst this data, we identified pathology serum lab tests as the richest source of data which could be positively linked with death and alive outcomes of patients. We initially examined risk ratios associated with death and survival in this subcohorts of patients in our first analysis which is displayed in Suppl. File 1. Next, we manually abbreviated this dataset and uploaded this as training data to build a decision-tree machine learning model in the platform 'Orange Data Mining'.^{9,13} We then assessed this model's performance against positively confirmed death and survival hospital outcomes (Suppl. File 2). Next, we compared this Tree model with two other AI-machine learning models, namely Neural Networks, and Random Forest and assessed the performance differences across the models (Suppl. File 3). Finally, using the same models we requested prediction outcomes using patient data, from a different hospital and geographic region, to assess whether the model's results were generalisable to other geographic regions and demographics (Suppl. File 4).^{4,14,15}

The anonymised electronic healthcare records of ICU patients with the diagnosis of infection was analysed from the open-source research group physio.net. Physio.net is a collaborative working group comprising over 200 hospitals from 140 countries and 50,000 registered researchers. Within this collection was a critical care database comprising 2,790 ICU patients with infection at Zigong Fourth People's

Hospital Sichuan, China from January 2019 to December 2020. ICU patient data is frequently closely monitored with high-granularity data and thus is appropriate for our research question.

Importantly, the Zigong Hospital database comprises extensive blood pathology data linked to patient outcomes. Since our scope of research specifically wanted to look at a subcohort of these patients that underwent surgery. We further filtered this database by International Classification of Disease (ICD) codes. These ICD codes are linked to each patient record. In our case we filtered patients to only those with 5,383 surgical-associated codes of all 35,686 ICD descriptions, this identified 2,124 unique patient ID numbers. Using Microsoft Excel, death and survival outcomes were then extracted for all these patient ID numbers. Laboratory tests for each of these patients with approximately ninety-five serological lab results (i.e. Creatine, Platelet, Urea) was classified into high, normal or low categories according to the pathology dataset's laboratory's healthy and pathological reference ranges. Frequency counts were then obtained for each group for each laboratory test (some patients had missing test values). The conversion of the quantitative lab values (i.e. 90,91,92) into ordinal values (i.e. low, normal, high) was necessary and reflective of standard pathological practice. The reason being that small quantitative differences in a patient's blood values (i.e. 90-100) may have significant clinical differences while (i.e. 80-90) may be normal. Therefore, serological data necessitated the differentiation into high, normal, low categorisation which would embed the homeostatic physiological normal ranges. From this reclassification we could compare four distinct sub-groups:

1. Survived with surgical intervention and high laboratory test value, 2) Deceased with surgical intervention with high laboratory test value 3) Survived with surgical intervention with normal or low specific laboratory test value 4) Deceased with surgical patients with normal or low specific laboratory marker value. Risk ratios were subsequently calculated from the categorised dataset to assess the association between laboratory markers and patient outcomes, and the resulting Risk Ratios were sorted. Standard errors for each laboratory finding and their ninety-five percent confidence intervals were calculated with the following formula. In general, higher frequency of observations narrowed confidence intervals, and rarer events increased standard error. The specific risk ratio and formula for standard error and confidence interval are provided in Supplementary File 6.

The Orange Data Mining (v3.40.0) application was utilized in this study to make a predictive model. A modified dataset with nine laboratory values for creatinine, potassium, urea, cholinesterase, eosinophil ratio, haematocrit, platelet, platelet haematocrit, and total protein, from a sample of 1,202 patients in a CSV file was imported as training data. Due to the nature of our database involving ordinal data,

we chose the Decision Tree Model as the first prototype to work with. Due to several limitations, not all 95 laboratory values were uploaded as training data, and we were restricted to nine features. Local RAM resource limitations and the large training set meant we were constrained in the number of features we could use. From a re-iterative troubleshooting process, we established that the standard workstation could train on ~10 features before an Out-of-Memory Error occurs in the application. We chose to limit the training to available local resources, and features found in 4.1, rather than establish a methodology that required more powerful servers but were unpracticable/unavailable in healthcare settings.

Within the Orange interface, these nine laboratory values were selected as features, and the outcome of death or survival was selected as the target. We requested the platform to ignore values with missing information, or irrelevant columns (i.e., missing or not missing, INP_N01).

We allocated ninety percent of patient data for training and allocated ten percent for prediction requests. The prediction request had outcome (death/alive) values manually removed for prediction requests, and we manually verified the accuracy of the prediction model after its predictions were provided. The snapshot of the workflow is shown below in Supplementary File 2. The widget 'CSV File Import' on the top left was used to upload our training data of 1081 patient records in the form of a CSV file. The widget 'CSV File Import (1)' shown on the middle bottom was used to upload our prediction request of 120 patients (outcomes removed).

We then investigated whether prediction outcomes could be improved by using other machine learning models. Two other models, Random Forest and Neural Networks were suitable for analysing ordinal data and were trained in the same manner as above. The Random Forest model is an improvement on the Decision Tree architecture by using the average of hundreds of different trees to reduce error, while the Neural Network model uses multiple layers of mathematical weights to find deep, non-linear patterns. A similar workflow to the above process was used; however, the 'data sampling' widget was utilised to select the split between ninety percent training and ten percent prediction and the 'test and score' widget which provided an automatic way to compare confusion matrices, F1, Precision and Recall values.

We then assess the model's performance against a separate demographic to ensure that there is not overfitting for patients in Zigong, China. To do this, we searched for a dataset that had similar ICU data and blood serology data and assessed the model's prediction performance. The hospital database from Beth Israel Deaconess Medical Center, Symile-MIMIC from Physio.net was chosen, as it featured lab data comprising fifty blood laboratory values recorded within twenty-four hours of admission. Despite this abundance of lab values, only six columns of the original nine training features (from the above methods) were

matching; nevertheless, we assessed whether our models could predict patient outcomes with these six lab features. Training data was modified to delete columns that were not present in our prediction file. The resultant workflow is shown as follows in Supplementary File 4. The data sampler widget was not compatible in this workflow as the datasets were from two different patient cohorts. Prediction results were saved and then true survival and mortality values were assigned correct and incorrect values. The accuracy of each of the three models against real survival data from Symile-MIMIC was then plotted and graphed.

To gain credentialed access to PhysioNet, registration followed by the completion of CITI Program ethics course was completed. A credentialing application was completed with a statement of research intent and the contact details of a research supervisor. Local Human Ethics Institutional approval was granted under identifier 2023/HE000829. A Data Use Agreement was signed to protect patient privacy and prohibit the sharing of data with unauthorized third parties. Original consent from patients at Zigong China Hospital was waived due to the retrospective design of the study and compliance with ethics approvals for a low-risk study with patient anonymisation. The original data collection was compliant with the Declaration of Helsinki and approved by the ethics committee of Zigong Fourth People's Hospital (Approval Number: 2020-065). Data from MIMIC were derived from MIMIC-IV, MIMIC-IV-ECG, and MIMIC-CXR-JPG databases. The Symile-MIMIC database was fully de-identified and obtained under a PhysioNet Credentialed Health Data Use Agreement. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2013.

Results

This project comprises four main results. Firstly, the manual analysis of patient data identified high risk blood values which can serve as biomarkers of sepsis and death. We then endeavoured to make a Decision Tree Machine Learning model which could receive and predict a patient's outcome prior to surgery. The performance of this Decision Model Tree model, however, was only slightly above average at fifty-six percent, whilst False Positives were forty-four percent (Figure 3). To determine if this result could be improved by a different mode, we trained two other ordinal data models (specifically Random Forest and Neural Networks). Coincidentally by doing this we also improved the Decision Tree's performance accuracy by two percent (fifty-eight percent) via the data sampling widget and a process called random seeding. The F1 performance of each model was similar with Neural Networks (0.625), Random Forest (0.650) and Decision Tree (0.663). Finally, we wanted to assess whether these models were generalisable to other demographic patients. We used a demographically distant

population (Symile-MIMIC database) to predict patient outcomes; this final result showed remarkably similar results with Decision Tree (fifty-eight percent), Neural Network (fifty-eight percent), Forest Model (fifty-eight percent) accuracy. Further details of each main result are described below.

Of ninety-two laboratory test values, fifty-four blood tests were statistically relevant to either an elevated risk of death

or were protective for patients. Forty-five lab tests were statistically significant in increasing Risk Ratios above 1. Nine lab tests were statistically significant in lowering risk of death (protective). While thirty-six blood values were not statistically relevant to death or survival. For concision, only the first ten lab values with the greatest Risk Ratios are shown in Figure 1, and all nine of the protective lab values are shown in Figure 1.

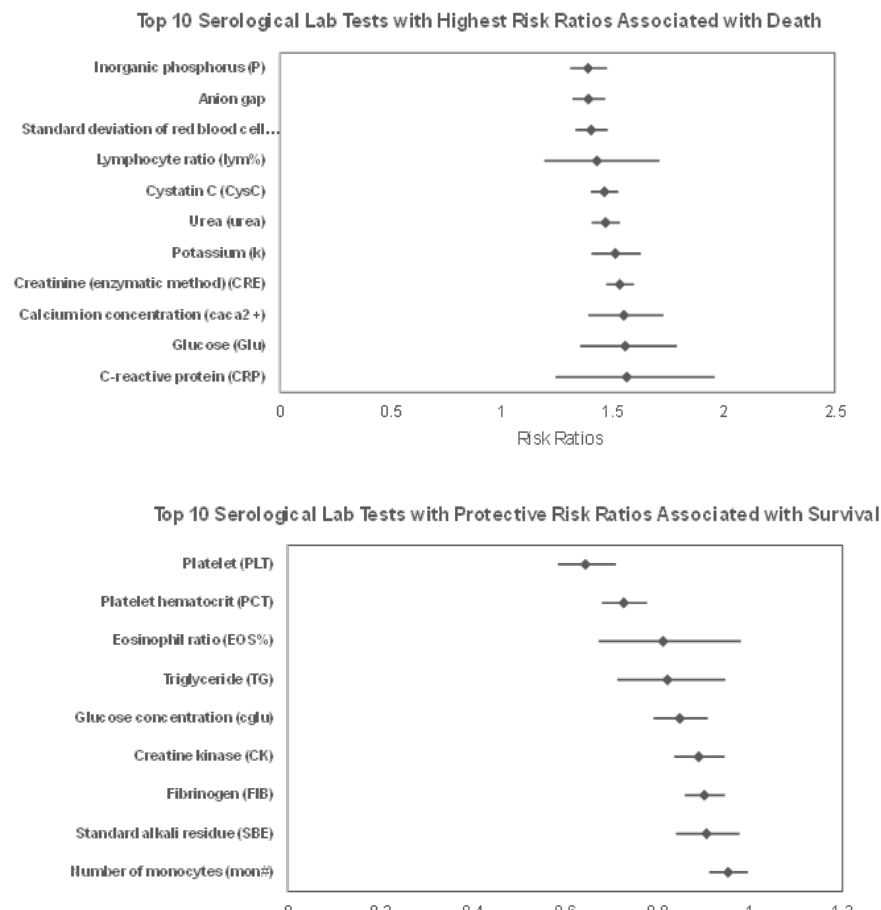


Figure 1: Laboratory test values associated with higher risks of mortality in ICU (top) and survival (bottom).

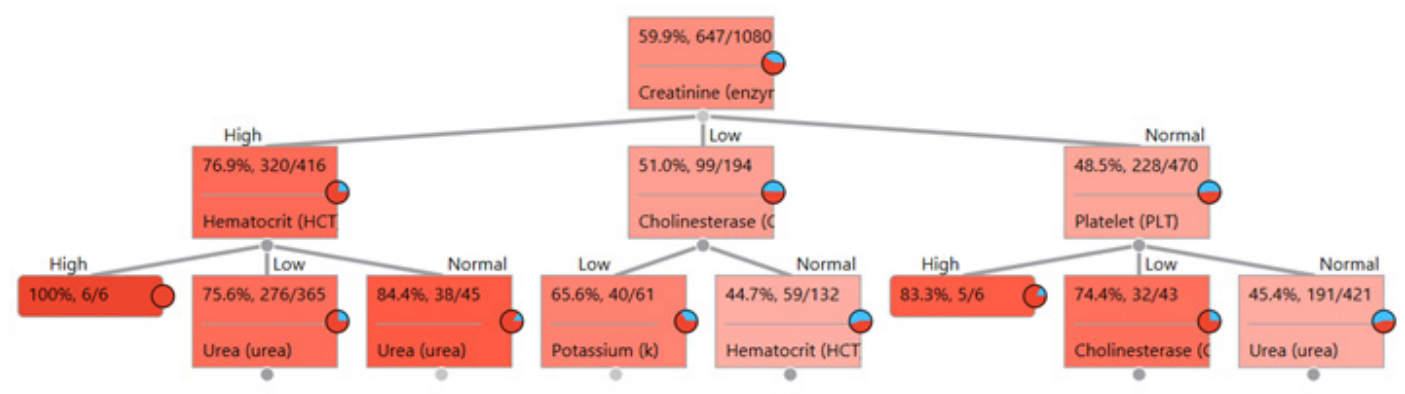


Figure 2: First three levels of generated decision tree on most important laboratory test values.

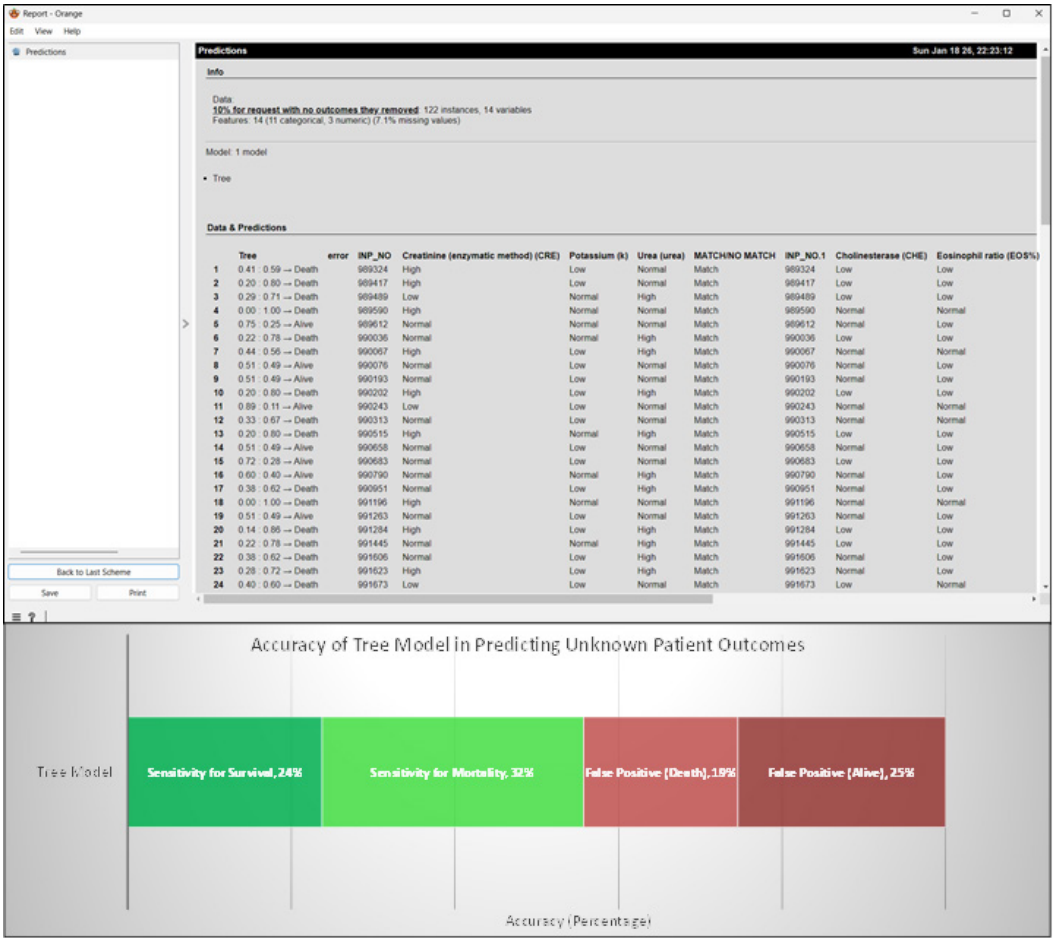


Figure 3: Output snapshot of decision tree prediction outcomes from Orange Data Mining platform (top) and accuracy of decision tree prediction outcomes compared against real patient outcomes(bottom).

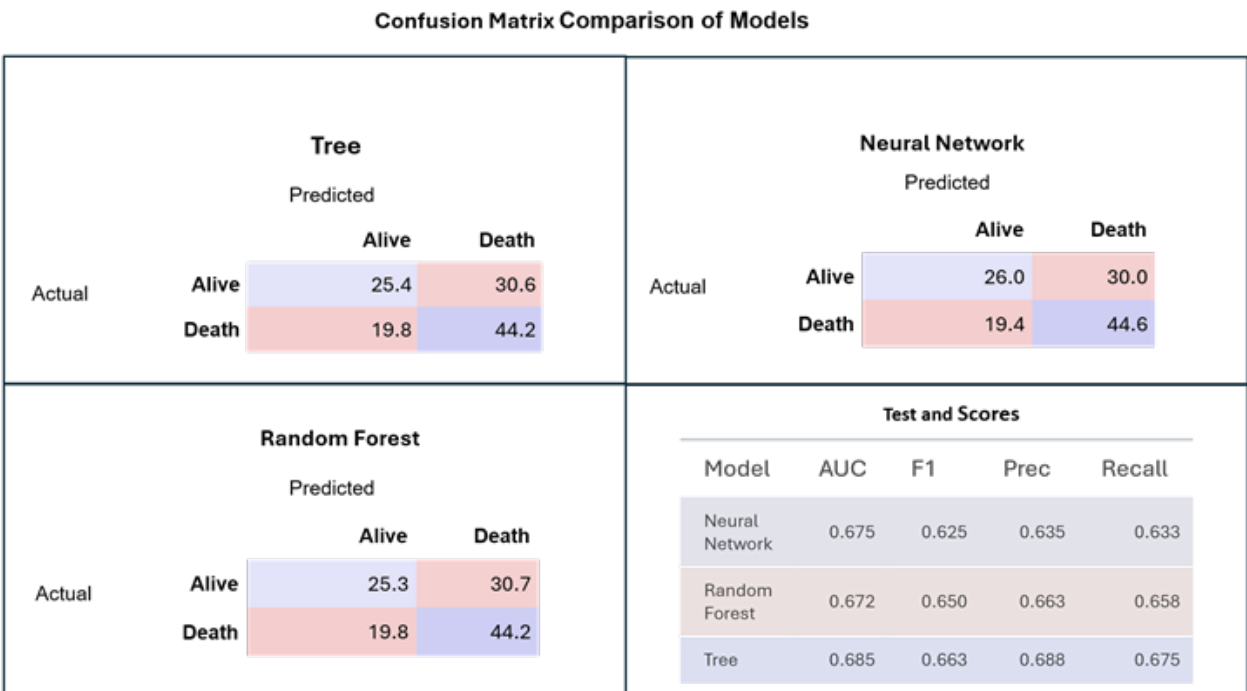


Figure 4: Performance differences between models, showing, AUC, confusion matrices and F1 scores.

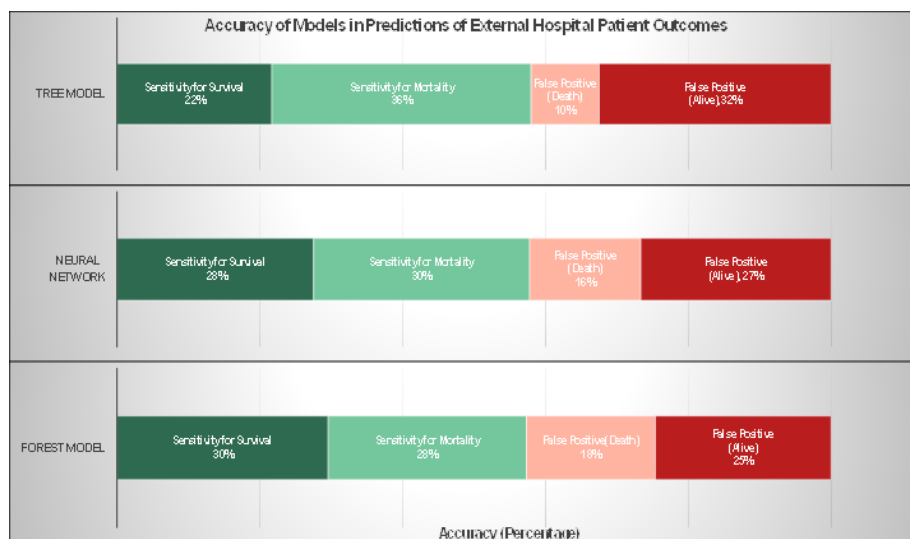
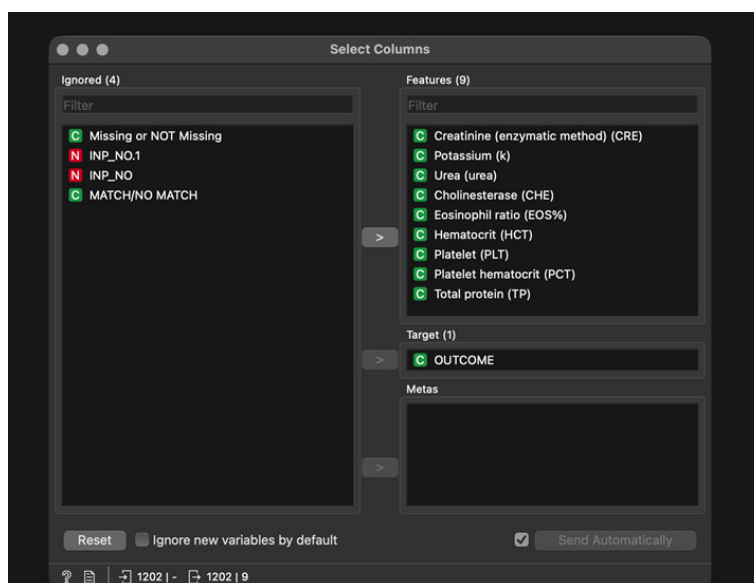
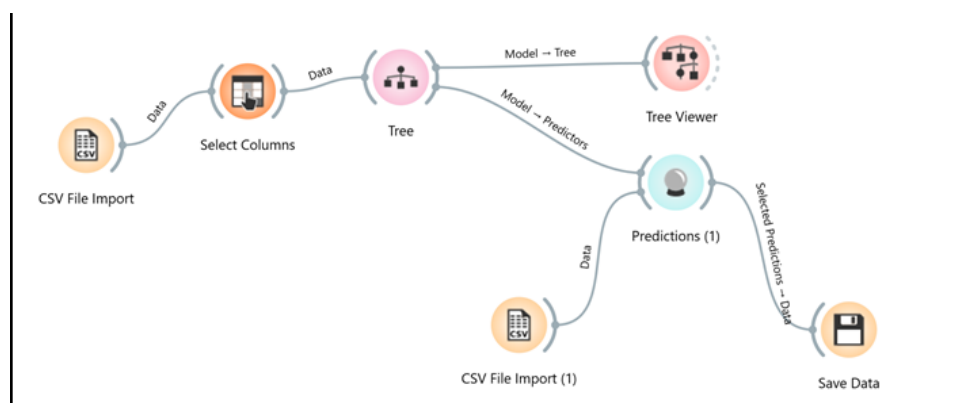


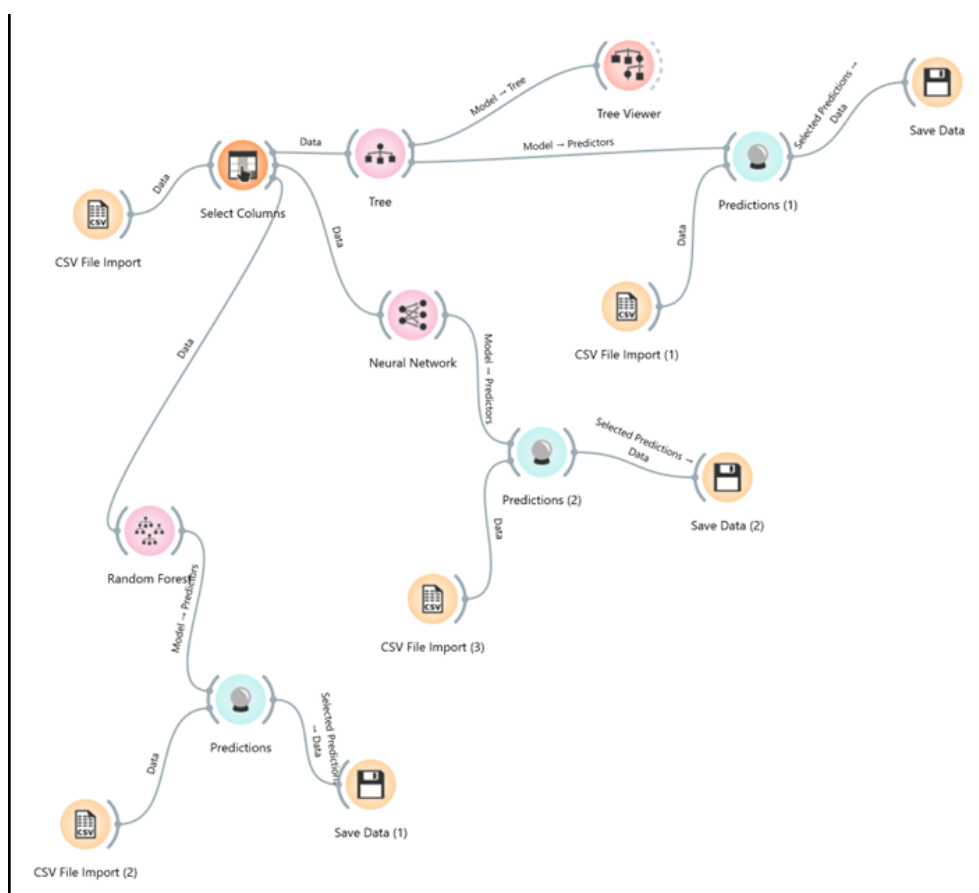
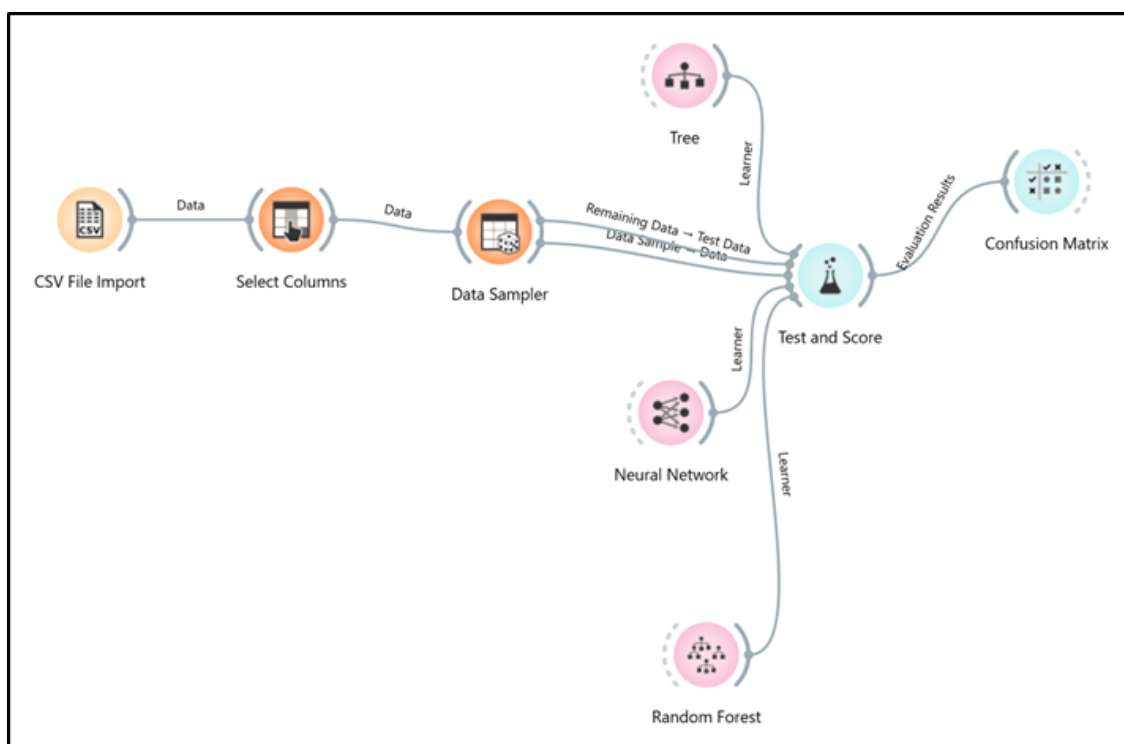
Figure 5: Accuracy of outcomes in Tree, Neural Network and Random Forest Models on Boston-Symile patient demographics.



Supplementary File 1: Feature and Outcome Selection in the Training of Decision Tree Model.



Supplementary File 2: Workflow Snapshot of Decision Tree Development.



Data & Predictions										
	Tree	error	INP_NO	OUTCOME	Hematocrit (HCT)	Platelet (PLT)	Creatinine (enzymatic method) (CRE)	Potassium (k)	Urea (urea)	Eosinophil
1	0.67 : 0.33 → Alive		10001176	?	Low	Normal	Normal	Normal	Normal	High
2	0.17 : 0.83 → Death		10001401	?	Low	High	Normal	Normal	High	High
3	0.67 : 0.33 → Alive		10001884	?	Normal	Normal	Normal	Normal	Normal	Low
4	0.67 : 0.33 → Alive		10002013	?	Low	Normal	Normal	Normal	Normal	Low
5	0.49 : 0.51 → Death		10002428	?	Low	Normal	Low	Low	Normal	Low
6	0.67 : 0.33 → Alive		10002428	?	Normal	Normal	Normal	Normal	Normal	Normal
7	0.26 : 0.74 → Death		10003400	?	Low	Low	Normal	Normal	Normal	Low
8	0.26 : 0.74 → Death		10003502	?	Low	Low	Normal	Normal	Normal	High
9	0.22 : 0.78 → Death		10004235	?	Low	Low	High	Normal	High	Low
10	0.67 : 0.33 → Alive		10007058	?	Low	Normal	Normal	Normal	Normal	Low

Data & Predictions										
	Neural Network	error	INP_NO	OUTCOME	Hematocrit (HCT)	Platelet (PLT)	Creatinine (enzymatic method) (CRE)	Potassium (k)	Urea (urea)	Eosinophil
1	0.71 : 0.29 → Alive		10001176	?	Low	Normal	Normal	Normal	Normal	High
2	0.83 : 0.17 → Alive		10001401	?	Low	High	Normal	Normal	High	High
3	0.67 : 0.33 → Alive		10001884	?	Normal	Normal	Normal	Normal	Normal	Low
4	0.69 : 0.31 → Alive		10002013	?	Low	Normal	Normal	Normal	Normal	Low
5	0.50 : 0.50 → Death		10002428	?	Low	Normal	Low	Low	Normal	Low
6	0.67 : 0.33 → Alive		10002428	?	Normal	Normal	Normal	Normal	Normal	Normal
7	0.27 : 0.73 → Death		10003400	?	Low	Low	Normal	Normal	Normal	Low
8	0.23 : 0.77 → Death		10003502	?	Low	Low	Normal	Normal	Normal	High
9	0.07 : 0.93 → Death		10004235	?	Low	Low	High	Normal	High	Low
10	0.69 : 0.31 → Alive		10007058	?	Low	Normal	Normal	Normal	Normal	Low

Data & Predictions										
	Random Forest	error	INP_NO	OUTCOME	Hematocrit (HCT)	Platelet (PLT)	Creatinine (enzymatic method) (CRE)	Potassium (k)	Urea (urea)	Eosinophil ratio
1	0.73 : 0.27 → Alive		10001176	?	Low	Normal	Normal	Normal	Normal	High
2	0.43 : 0.57 → Death		10001401	?	Low	High	Normal	Normal	High	High
3	0.69 : 0.31 → Alive		10001884	?	Normal	Normal	Normal	Normal	Normal	Low
4	0.73 : 0.27 → Alive		10002013	?	Low	Normal	Normal	Normal	Normal	Low
5	0.58 : 0.42 → Alive		10002428	?	Low	Normal	Low	Low	Normal	Low
6	0.83 : 0.17 → Alive		10002428	?	Normal	Normal	Normal	Normal	Normal	Normal
7	0.43 : 0.57 → Death		10003400	?	Low	Low	Normal	Normal	Normal	Low
8	0.43 : 0.57 → Death		10003502	?	Low	Low	Normal	Normal	Normal	High
9	0.02 : 0.98 → Death		10004235	?	Low	Low	High	Normal	High	Low
10	0.73 : 0.27 → Alive		10007058	?	Low	Normal	Normal	Normal	Normal	Low

Supplementary File 5: Snapshot of first ten prediction outcomes in Tree, Neural Network and Random Forest Models.

Calculate the Standard Error (SE) of the Log RR

$$SE_{\ln RR} = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}}$$

- **Position a and c :** The number of people who died in each group.
- **Position n_1 and n_2 :** The total number of people in each group.

Calculate the 95% Confidence Interval (The Range)

- **Lower Bound** = $e^{\ln(RR) - (1.96 \times SE_{\ln RR})}$
- **Upper Bound** = $e^{\ln(RR) + (1.96 \times SE_{\ln RR})}$

Supplementary File 6: Formulas used for Risk Ratio calculations.

A Decision Tree learning model was successfully generated with the training data comprising a 9 levels depth, 115 nodes, and 61 leaves. Each node delineates a data split based on feature values, while leaves representing terminal nodes provide final outcome prediction. For brevity, only the first three levels of the visualised decision tree are shown in Figure 2. An abbreviated snapshot of 24 of the 120 prediction outcomes are shown below. The second column describes two probability scores for alive and death respectively separated by a ratio symbol “:”, right-sided probabilities over 0.5 are treated as a prediction for a death outcome, whereas a left-sided probabilities over 0.5 is considered a prediction for an alive outcome (Figure 3). All one-hundred-twenty results were fully tabulated and compared against real outcomes in Figure 3.

When Decision Tree outcomes were saved and compared against the original real patient outcomes, only fifty-six percent of outcomes were correctly predicted. Correct predictions comprised correct predictions for survival twenty-four percent, correct predictions for mortality thirty-two percent, incorrect predictions for death nineteen percent, and incorrect predictions for survival twenty-five percent.

Discussion

The introduction of the random sampler widget in the workflow appears to have improved training and prediction outcomes by an effect called random seeding, as opposed to the fixed seeding initially used in 5.2. Random seeding

randomly selects ninety percent of patient data for training and randomly selects ten percent for prediction requests. This result appears to improve decision tree prediction outcomes by two percent. Development of two other training models, Neural Networks and Random Forests shows similar prediction accuracy rates with Neural Networks (fifty-eight percent) and Random Forest (fifty-seven point nine percent). An averaged F1 score (averaged target class Alive and Death) was also automatically generated for each model. The F1 score is a more robust metric for an imbalanced dataset with higher survivals, like ours. Overall, the alternative models (Neural Networks and Random Forests) have not significantly improved prediction accuracy, and F1 scores declined.

One would expect large differences due to the geographic and demographic differences between patient data in Zigong, China, and Boston, Massachusetts; however accuracy rates were remarkably similar. For brevity, the snapshot of the first ten Prediction Outputs of each model is shown (Suppl. File 11). The Boston-Symile-MIMIC-database represents a much larger patient database with 11622 patient records. This workflow did not require data sampling or a training/testing split, as all training was performed on the Zigong China database and all prediction requests were performed on the Boston-Symile-MIMIC database. All 11622 outcomes were then compared against true patient outcomes, and their accuracy was plotted on a stacked bar chart (Figure 5).

All 11622 prediction outcomes were compared against original true patient outcomes, and their accuracy was plotted on a stacked bar chart in Figure 5. Overall accuracy (Sensitivity Survival + Sensitivity for Mortality) of the Decision Tree model on Boston patients was fifty-eight percent, a plus two percent improvement over the original model, but the same as when the data sampling widget was integrated in the third part (5.3) of our study. Similarly, the Neural Network model and the Forest model's accuracy float at roughly fifty-eight percent. The largest difference can be seen in the False Positive Death and False Positive Alive outcomes across the three models. Whilst sensitivity for survival and mortality is self-explanatorily important, the clinical significance of False Positives is worth exploring. Understanding that a patient could be at a true high risk of mortality could influence medical interventions and treatments (i.e., being more conservative with surgery or increasing monitoring). However, a False Positive for Death (when the patient actually survives) may also mean that beneficial surgical interventions are withheld, or similarly excessive medical resources may be allocated to this patient when they have a low risk of dying. Nevertheless, from a purely risk-avoidance perspective, a model like the Forest Model with Sensitivity + False Positives (Death) = seventy-six percent may be a more useful model than the Decision Tree model. If all deaths are trying to be avoided, regardless of resource availability.

The persistence of false positive classifications in the study

highlights the critical trade-off between sensitivity and specificity in mortality prediction. False positive predictions (of death) may lead to unnecessary resource utilisation in the context of staffing, cost of technology, and diagnostic resources. However, prioritising sensitivity may still be justified in high-risk surgical settings where consequences of missed mortality are substantial.

Other future improvements are likely to come through licensing and confidentiality agreements with larger compute services and consensus in health care data regulations. As described earlier, there are significant legal and ethics hurdles preventing the use of hospital data sets in noncompliant browser AI models or larger-scale online AI models; however, with future developments and inevitable API agreements established, we anticipate significant generational changes in the quality and utility of these clinical tools.^{17,18}

Comparisons of our ML model's performances against the literature on Acute Physiology and Chronic Health Evaluation II (APACHE II) and Sequential Organ Failure Assessment (SOFA) score shows major differences. Whilst some individual APACHE II studies have reported excellent sensitivity rates of 92.8% for detection of non-survivors¹⁹, other systematic reviews have indicated that AI models have demonstrated higher AUCs (0.82-0.90) compared to APACHE II (0.70-0.78).²⁰ Whereas, our study has explored a niche subgroup of the dataset, namely post-surgery patients, our results have not been able to replicate these higher sensitivity and specificity rates.

Limitations of our study

This study was an exploration of whether machine learning models could be used to predict patient outcomes. Whilst our initial exploration of Risk-Ratios found a number of useful risk factors, the ordinal data prediction models have an unusually low accuracy level (fifty-eight percent), F1 (0.625-0.663), which may preclude their use in the clinical setting.

Paradoxically, while advanced data scientists may have reasoning models nearing ninety-five to ninety-nine percent in their prediction models, we speculate a number of reasons for why we observed this disparity. Firstly, our choice of platform, Orange Data Mining, is compliant for ethical data analysis on local computer systems (not sent to a third party). Conversely, the international trend is to move all compute into GPU-accelerated instances hired from the Google Cloud Platform or Amazon Web Services. This meant that we were unable to utilise all ninety-five laboratory test values in our training data to better develop our model.

Secondly, the authors of this study are training health-care clinicians, we were limited by software coding expertise and therefore used the no-coding application Orange Data Mining. A more unique coding architecture for our model may have produced better prediction outcomes.

Given that our target audience is clinical practitioners, we sought to minimize technical barriers and opportunities for replication.

Conclusion

Thus far, the accuracies of our machine learning models were modest, with performance ranging from fifty-six to fifty eight percent. While these were statistically robust results, their lower accuracies preclude their use in the clinic until higher accuracy is achieved. Interestingly, the consistency across multiple modelling approaches suggests that

outcome prediction may be constrained by the size of our dataset rather than model choice alone. In future studies, the utilization of commercial compute services like Google Cloud Compute or Amazon Web Services can extend feature size selection and therefore produce higher fidelity in prediction performance.

Nevertheless, despite this lower accuracy rate, when our models are trained on the Chinese ICU cohort and then applied to a North American population, a very similar performance was observed.

Integrity and compliance statements

Ethics approval

Approved by research integrity and ethics administration of university of Sydney on 19th July 2024(Ref: 2023/HE000829), and this study was conducted at Griffith University.

Informed consent

This study used secondary data and it was not applicable

Source of Funding

No external funding was received for this work.

Conflict of Interest

The authors declare no competing interests.

Data availability

No datasets were generated or analyzed for this study.

Provenance and Peer Review

Not commissioned; externally peer reviewed. Double-blind peer review.

AI/LLM disclosure

No AI-assisted tools were used in this manuscript.

License

© The Author(s) 2026. This work is licensed under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Acknowledgements

The authors would like to thank all the study participants of the study.

Author Contributions (CRediT)

John Ng: Design, concept, data collection, data analysis, manuscript preparation, manuscript review and editing.

David Lee: Design, concept, data collection, Literature review, manuscript review and edition

Sneha Chhetri: Literature Search, Data Curation (supporting), Methodology (supporting), Manuscript draft, Editing (Supporting).

Bios

John Ng is a PhD qualified scientist on extended leave from a Research Masters at the University of Sydney, and is currently enrolled as a medical student at the Griffith University Medical Program. He is dedicated to advancing clinical technology in the clinic.

David Lee is medical student at School of Medicine and Dentistry, Griffith University. He is interested in trauma surgery and emergency medicine.

Sneha Chhetri is a MD candidate at the school of medicine and dentistry, Griffith University. She is interested in utilising health data to understand trends and facilitate safe clinical care.

References

- Moreno AP, Alvarez AA, Conde SM, Flecha MP, Ureña MG. The intensive care unit in the postoperative period of major abdominal surgery. *Medicina Intensiva (English Edition)*. 2019;43(9):569-77. DOI: [10.1016/j.medine.2019.05.011](https://doi.org/10.1016/j.medine.2019.05.011)
- Critical care database comprising patients with infection at Zigong Fourth People's Hospital (version 1.1) [Internet]. 2022. DOI: [10.13026/xpt9-z726](https://doi.org/10.13026/xpt9-z726)
- Xu P, Chen L, Zhu Y, Yu S, Chen R, Huang W, et al. Critical Care Database Comprising Patients With Infection. *Frontiers in Public Health*. 2022;Volume 10 - 2022. DOI: [10.3389/fpubh.2022.852410](https://doi.org/10.3389/fpubh.2022.852410) PMID: 35372245 PMCID: PMC8968758
- Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):e215-e20. DOI: [10.1161/01.CIR.101.23.e215](https://doi.org/10.1161/01.CIR.101.23.e215)
- Beyene RT, Derryberry SL, Barbul A. The effect of comorbidities on wound healing. *Surgical Clinics*. 2020;100(4):695-705. DOI: [10.1016/j.suc.2020.05.002](https://doi.org/10.1016/j.suc.2020.05.002) PMID: 32681870 PMCID: PMC10586696
- Fan S-L, Miller NS, Lee J, Remick DG. Diagnosing sepsis-The role of laboratory medicine. *Clinica chimica acta*. 2016;460:203-10. DOI: [10.1016/j.cca.2016.07.002](https://doi.org/10.1016/j.cca.2016.07.002) PMID: 27387712 PMCID: PMC4980259
- Van den Berg M, Van Beuningen F, Ter Maaten J, Bouma H. Hospital-related costs of sepsis around the world: A systematic review exploring the economic burden of sepsis. *Journal of critical care*. 2022;71:154096. DOI: [10.1016/j.jcrc.2022.154096](https://doi.org/10.1016/j.jcrc.2022.154096) PMID: 35839604

8. Back JS, Jin Y, Jin T, Lee SM. Development and validation of an automated sepsis risk assessment system. *Research in nursing & health*. 2016;39(5):317-27. DOI: [10.1002/nur.21734](https://doi.org/10.1002/nur.21734) PMID: 27327444
9. Wong HR, Lindsell CJ, Pettilä V, Meyer NJ, Thair SA, Karlsson S, et al. A Multibiomarker-Based Outcome Risk Stratification Model for Adult Septic Shock*. *Critical Care Medicine*. 2014;42(4):781-9. DOI: [10.1097/CCM.0000000000000106](https://doi.org/10.1097/CCM.0000000000000106) PMID: 24335447 PMCID: PMC4620515
10. Makam AN, Nguyen OK, Auerbach AD. Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: a systematic review. *Journal of hospital medicine*. 2015;10(6):396-402. DOI: [10.1002/jhm.2347](https://doi.org/10.1002/jhm.2347) PMID: 25758641 PMCID: PMC4477829
11. De Oliveira H, Prodel M, Lamarsalle L, Inada-Kim M, Ajayi K, Wilkins J, et al. "Bow-tie" optimal pathway discovery analysis of sepsis hospital admissions using the Hospital Episode Statistics database in England. *JAMIA open*. 2020;3(3):439-48. DOI: [10.1093/jamiaopen/ooaa039](https://doi.org/10.1093/jamiaopen/ooaa039) PMID: 33215077 PMCID: PMC7660952
12. Morrill J, Kormilitzin A, Nevado-Holgado A, Swaminathan S, Howison S, Lyons T, editors. The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. 2019 *Computing in Cardiology (CinC)*; 2019: IEEE. DOI: [10.22489/CinC.2019.014](https://doi.org/10.22489/CinC.2019.014)
13. Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, et al. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*. 2013;14(Aug):2349-53.
14. Saporta A, Puli AM, Goldstein M, Ranganath R. Symile-MIMIC: a multimodal clinical dataset of chest X-rays, electrocardiograms, and blood labs from MIMIC-IV (version 1.0.0). *PhysioNet*. 2025.
15. Saporta A, Puli AM, Goldstein M, Ranganath R. Contrasting with Symile: Simple Model-Agnostic Representation Learning for Unlimited Modalities. *Advances in Neural Information Processing Systems*. 2024;37:56919-57. DOI: [10.52202/079017-1814](https://doi.org/10.52202/079017-1814)
16. Bates DW, Goldman L, Lee TH. Contaminant blood cultures and resource utilization: the true consequences of false-positive results. *Jama*. 1991;265(3):365-9. DOI: [10.1001/jama.1991.03460030071031](https://doi.org/10.1001/jama.1991.03460030071031) PMID: 1984535
17. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare: Elsevier*; 2020. p. 295-336. DOI: [10.1016/B978-0-12-818438-7.00012-5](https://doi.org/10.1016/B978-0-12-818438-7.00012-5) PMCID: PMC7332220
18. Whicher D, Ahmed M, Israni ST. Health care artificial intelligence: law, regulation, and policy. *Artificial intelligence in health care: the hope, the hype, the promise, the peril*. Washington (DC): National Academies Press
19. TianY, YaoY, ZhouJ, DiaoX, ChenH, CaiK, et al. DynamicAPACHEII score to predict the outcome of intensive care unit patients. *Front Med*. 2022;8. DOI: [10.3389/fmed.2021.744907](https://doi.org/10.3389/fmed.2021.744907) PMID: 35155461 PMCID: PMC8826444
20. Pradhan V, Shekhar H, Munda PK, Tiwari AK, Jha S, Rai P. Accuracy of artificial intelligencebased models versus traditional scoring systems (APACHE, SOFA, SAPS) for predicting mortality in ICU patients: a systematic review and metaanalysis. *medRxiv*. 2026;2026.01.14.26344000. DOI: [10.64898/2026.01.14.26344000](https://doi.org/10.64898/2026.01.14.26344000)