



# LUND UNIVERSITY

## ICU prognostication: Time to re-evaluate? Register-based studies on improving prognostication for patients admitted to the intensive care unit (ICU)

Andersson, Peder

2021

*Document Version:*

Publisher's PDF, also known as Version of record

[Link to publication](#)

*Citation for published version (APA):*

Andersson, P. (2021). *ICU prognostication: Time to re-evaluate? Register-based studies on improving prognostication for patients admitted to the intensive care unit (ICU)*. Lund University, Faculty of Medicine.

*Total number of authors:*

1

### General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

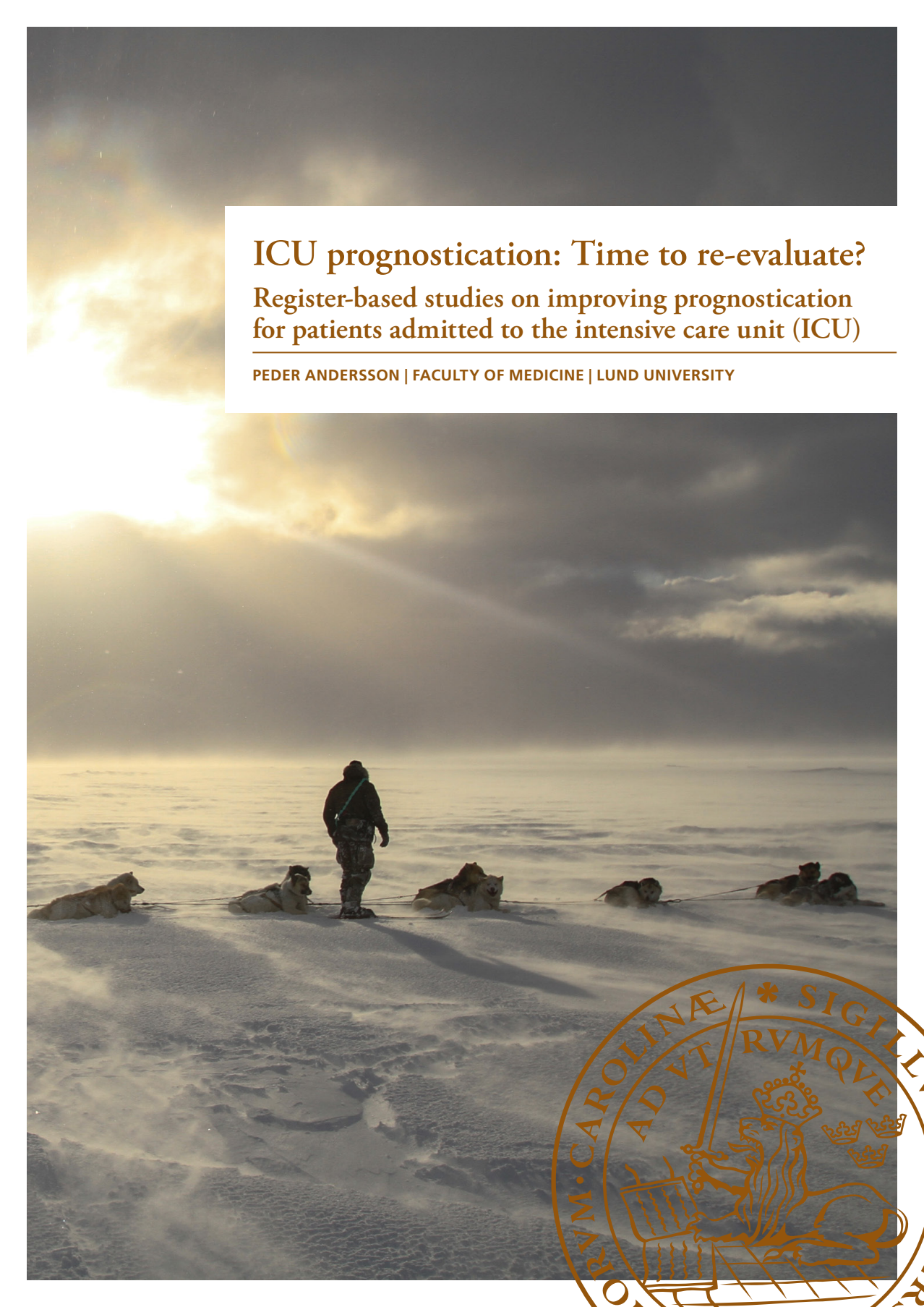
Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117  
221 00 Lund  
+46 46-222 00 00




# ICU prognostication: Time to re-evaluate?

## Register-based studies on improving prognostication for patients admitted to the intensive care unit (ICU)

PEDER ANDERSSON | FACULTY OF MEDICINE | LUND UNIVERSITY





The scoring systems for critically ill patients were developed almost 40 years ago and have been updated several times since. It is essential that these scoring systems perform at the absolute highest level, and they must, therefore, be continuously re-evaluated and challenged to improve their performance. By testing promising variables and new methods, we can continue to secure the best possible prediction models for our critically ill patients overall and for specific diagnoses.



**FACULTY OF  
MEDICINE**

Department of Clinical Sciences Lund  
Section of Anesthesiology and Intensive Care

Lund University, Faculty of Medicine  
Doctoral Dissertation Series 2021:11  
ISBN 978-91-8021-017-1  
ISSN 1652-8220



‘Logic will get you from A to Z;  
imagination will get you everywhere.’

– Albert Einstein



# ICU prognostication: Time to re-evaluate?

Register-based studies on improving prognostication for  
patients admitted to the intensive care unit (ICU)

Peder Andersson



**LUND**  
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden.  
To be defended at Belfragesalen, BMC, Klinikgatan 32, Lund.  
Thursday, the 11<sup>th</sup> of February, 2021 at 1 p.m.

*Faculty opponent*

Christian Fynbo Christiansen

*Supervisor*

Attila Frigyesi

*Co-supervisor*

Jonas Björk

<b>Organisation</b> LUND UNIVERSITY Department of Clinical Sciences, Anaesthesiology and Intensive Care Author Peder Andersson		<b>Document name</b> DOCTORAL DISSERTATION
		<b>Date of issue</b> February 11 <sup>th</sup> 2021
		Sponsoring organisation
<b>Title and subtitle:</b> ICU prognostication: Time to re-evaluate? Register-based studies on improving prognostication for patients admitted to the intensive care unit (ICU)		
<b>Abstract</b> <b>Background:</b> ICU prognostication is difficult because of patients' prior comorbidities and their varied reasons for admission. The model used for ICU prognostication in Sweden is the Simplified Acute Physiology Score 3 (SAPS 3), which uses information gathered within one hour of ICU admission to predict 30-day mortality. Since the SAPS 3 model was introduced, no biomarkers have been added to it to improve its prognostic performance. For comatose patients admitted to the ICU after cardiac arrest, the prognostication performed after 72 h will either result in the continued observation of the patient or the withdrawal of life-sustaining treatment. <b>Purpose:</b> 1) To investigate whether adding the biomarker lactate (study I) or high-sensitivity troponin T (hsTnT) (study II) to SAPS 3 adds prognostic value. 2) To investigate whether using a supervised machine learning algorithm called artificial neural networks (ANNs) can improve the prognostic performance of SAPS 3 (study III). 3) To explore whether ANNs can create reliable predictions for comatose patients at the time of hospital admission (study IV) and during the first three days after ICU admission, with or without promising biomarkers (study V). <b>Methods:</b> 1) To investigate whether the laboratory values of lactate or hsTnT could improve the performance of SAPS 3, we combined patients' laboratory values on ICU admission at Skåne University Hospital with their SAPS 3 score. 2) Based on all first-time ICU admissions in Sweden from 2009–2017 as retrieved from the Swedish Intensive Care Registry (SIR), we investigated whether ANNs could improve SAPS 3 using the same variables. 3) All out-of-hospital cardiac arrest (OHCA) patients from the Target Temperature Management trial were included for data analysis. Background and prehospital data, along with clinical variables at admission, were used in study IV. Clinical variables from the first three days were used in study V along with different levels of biomarkers defined as clinically accessible (e.g. neuron-specific enolase, or NSE) and research-grade biomarkers (e.g. neurofilament light, or NFL). Patient outcome was the dichotomised Cerebral Performance Category scale (CPC); a CPC of 1–2 was considered a good outcome, and a CPC of 3–5 was considered a poor outcome. <b>Results:</b> 1) Both lactate and hsTnT were independent SAPS 3 predictors for 30-day mortality in the logistic regression model. In a subgroup analysis, the use of lactate improved the area under the receiver operating characteristic curve (AUROC) for cardiac arrest and septic patients, and the use of hsTnT improved the AUROC for septic patients. 2) The overall performance of the SAPS 3 model in Sweden was improved by the use of ANNs. Both the discrimination (AUROC 0.89 vs 0.85, $p < 0.001$ ) and the calibration were improved when the two models were compared on a separate test set ( $n = 36,214$ ). 3) An ANN model outperformed a logistic-regression-based model in predicting poor outcome on hospital admission for OHCA patients. Incorporating biomarkers such as NSE improved the AUROC over the course of the first three days of the ICU stay; when NFL was incorporated, the prognostic performance was excellent from day 1. <b>Conclusion:</b> Lactate and hsTnT probably add prognostic value to SAPS 3 for patients admitted to the ICU with sepsis or after cardiac arrest (lactate only). An ANN model was found to be superior to the SAPS 3 model (Swedish modification) and corrected better for age than SAPS 3. A simplified ANN model with eight variables showed performance similar to that of the SAPS 3 model. For comatose OHCA patients, an ANN model improved the accuracy of the prediction of the long-term neurological outcome at hospital admission. Furthermore, when it used cumulative information from the first three days after ICU admission, an ANN model showed promising prognostic performance on day 3 when it incorporated clinically accessible biomarkers such as NSE, and it showed promising performance on days 1–3 when it incorporated research-grade biomarkers such as NFL.		
<b>Keywords:</b> Intensive care, critical care, mortality, Cerebral Performance Category, out-of-hospital cardiac arrest, cardiac arrest, sepsis, lactate, troponin, neuron-specific enolase, neurofilament light, age, elderly, prognostication, prediction, scoring system, neural net, artificial neural network, deep learning, artificial intelligence		
Classification system and/or index terms (if any)		
Supplementary bibliographical information		<b>Language</b> English
ISSN and key title 1652-8220		<b>ISBN</b> 978-91-8021-017-1
Recipient's notes	<b>Number of pages</b> 84	Price
	Security classification	

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature



Date 2021-01-11

# ICU prognostication: Time to re-evaluate?

Register-based studies on improving prognostication for  
patients admitted to the intensive care unit (ICU)

Peder Andersson



**LUND**  
UNIVERSITY



Cover: Northeast Greenland National Park ©Michael Andersson, a former member of the Sirius Sledge Patrol, part of the Danish special forces.

Copyright pp 1-84 Peder Andersson

Paper 1 © Wiley

Paper 2 © Elsevier

Paper 3 © The Authors (Open Access). Published by BMC.

Paper 4 © The Authors (Open Access). Published by BMC.

Paper 5 © The Authors (Unpublished manuscript)

Faculty of Medicine  
Department of Clinical Sciences, Lund  
Section of Anaesthesiology and Intensive Care

ISBN 978-91-8021-017-1

ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University, Lund 2021



Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at [www.mediatryck.lu.se](http://www.mediatryck.lu.se)

**MADE IN SWEDEN** 

*To my family*



# Table of Contents

List of publications .....	11
Abbreviations .....	12
<b>Background .....</b>	<b>15</b>
The intensive care unit .....	15
History .....	15
Modern intensive care medicine.....	16
Intensive care medicine in Sweden .....	16
ICU prognostication .....	17
Background .....	17
ICU prognostication in Sweden .....	18
Improving ICU prognostication .....	20
Recalibration of current models .....	20
Adding single predictors .....	20
Machine learning .....	22
Diagnosis-specific predictions .....	23
Post-cardiac arrest syndrome and prognostication .....	23
Post-cardiac arrest syndrome.....	24
Post-cardiac arrest prognostication.....	24
Biomarkers in cardiac arrest prognostication .....	27
<b>Aims of the thesis .....</b>	<b>29</b>
<b>Methods and materials.....</b>	<b>31</b>
Sources of data .....	32
PASIVA .....	32
Swedish Intensive Care Registry.....	32
The Target Temperature Management trial.....	32
Methods.....	34
Performance measures.....	34
Artificial neural network (ANN) – A brief introduction .....	37
Value of additional biomarkers (studies I and II).....	39
Using ANNs to short-term predict mortality (study III) .....	40

Predicting neurological outcome after out-of-hospital cardiac arrest (studies IV and V) .....	41
Software .....	44
Ethics.....	44
<b>Results.....</b>	<b>45</b>
Value of additional biomarkers (studies I and II).....	45
Study I – The prognostic value of lactate.....	45
Study II – The prognostic value of high-sensitivity troponin T .....	46
Using ANNs to predict short-term mortality (study III) .....	48
Predicting neurological outcome after out-of-hospital cardiac arrest (studies IV and V) .....	51
Study IV – Cardiac arrest prognostication on admission .....	51
Study V – Cardiac arrest prognostication during the first three days after ICU admission .....	54
<b>Discussion .....</b>	<b>59</b>
Value of additional biomarkers (studies I and II).....	60
Using ANNs to predict short-term mortality (study III) .....	61
Predicting neurological outcome after out-of-hospital cardiac arrest (studies IV and V) .....	62
Improving ICU prognostication .....	63
<b>Conclusions .....</b>	<b>65</b>
<b>Future perspectives .....</b>	<b>67</b>
Continuing to improve early ICU prognostication.....	67
Developing dynamic models .....	67
Morbidity prognostication.....	67
Prediction on an individual level.....	68
Improving post–cardiac arrest prognostication .....	68
<b>Populärvetenskaplig sammanfatning .....</b>	<b>69</b>
<b>Acknowledgements and grants.....</b>	<b>71</b>
<b>References .....</b>	<b>75</b>

# List of publications

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals.

- I. **Andersson P**, Frigyesi A. Lactate improves SAPS 3 prognostication. *Acta Anaesthesiol Scand*. 2018;62(2):220–5.
- II. **Andersson P**, Frigyesi A. High-sensitivity troponin T is an important independent predictor in addition to the Simplified Acute Physiology Score for short-term ICU mortality, particularly in patients with sepsis. *J Crit Care*. 2019;53:218–22.
- III. Holmgren G, **Andersson P**, Jakobsson A, 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. *J intensive care*. 2019;7:44.
- IV. Johnsson J, Björnsson O, **Andersson P**, Jakobsson A, Cronberg T, Lilja G, Friberg H, Hassager C, Wise M, Nielsen N\*, Frigyesi A\*. Artificial neural networks improve early outcome prediction and risk classification in out-of-hospital cardiac arrest patients admitted to intensive care. *Crit Care*. 2020;24(1):474.
- V. **Andersson P**, Johnsson J, Björnsson O, Cronberg T, Hassager C, Zetterberg H, Stammet P, Undén J, Kjaergaard J, Friberg H, Blennow K, Lilja G, Wise M, Dankiewicz J, Nielsen N\*, Frigyesi A\*. Predicting neurological outcome after out-of-hospital cardiac arrest with cumulative information; development and internal validation of an artificial neural network algorithm. Manuscript submitted to *Crit Care*.

Papers I and II are reprinted with permission of the copyright owners. Papers III and IV are open access. \* Niklas Nielsen and Attila Frigyesi contributed equally (papers IV and V).

## Abbreviations

AI	Artificial intelligence
AMI	Acute myocardial infarction
ANN	Artificial neural network
APACHE	Acute Physiology And Chronic Health Evaluation
AUROC	Area under the receiver operating characteristic curve
BNP	N-terminal pro-B-type natriuretic peptide
CI	Confidence interval
CPC	Cerebral Performance Category scale
CPR	Cardiopulmonary resuscitation
CT	Computed tomography
ECG	Electrocardiogram
ECMO	Extracorporeal membrane oxygenation
EEG	Electroencephalogram
EMR	Estimated mortality rate
ERC	European Resuscitation Council
ESICM	European Society of Intensive Care Medicine
FNR	False-negative rate
FPR	False-positive rate
GCS	Glasgow Coma Scale
GFAP	Glial fibrillary acidic protein
GiViTI	Italian Group for the Evaluation of Interventions in Intensive Care Medicine
hsTnT	High-sensitivity troponin T
ICD-10	10th revision of the International Classification of Diseases
ICU	Intensive care unit
IL-6	Interleukin-6
MPM	Mortality Probability Model
MRI	Magnetic resonance imaging
mRS	Modified Rankin Scale
NFL	Neurofilament light
NSE	Neuron-specific enolase
OHCA	Out-of-hospital cardiac arrest
OMR	Observed mortality rate

PASIVA	'Patientadministrativt system för Intensivvårdsavdelningar'
PCAS	Post-cardiac arrest syndrome
PCT	Procalcitonin
PIM3	Paediatric Index of Mortality 3
RLS	Reaction level scale
ROC	Receiver operating characteristic
ROSC	Return of spontaneous circulation
SAPS 3	Simplified Acute Physiology Score 3
SHAP	Shapley additive explanations
SIR	The Swedish Intensive Care Registry
SML	Supervised machine learning
SMR	Standardised mortality ratio
SOFA	Sequential Organ Failure Dysfunction score
SSEP	Short-latency somatosensory evoked potentials
S100B	S100 calcium-binding protein B
Tau	Tau protein
TnT	Troponin T
TPR	True-positive rate
TTM	Targeted temperature management
TTM-trial	Target temperature management trial
UCHL1	Ubiquitin carboxy-terminal hydrolase L1
WLST	Withdrawal of life-sustaining therapy





# Background

## The intensive care unit

### History

Intensive care medicine's true origin is difficult to distinguish. It is a continuum of events: different contributions from pioneers, medical and technological advancements, combined with the constant demand to cure or relieve human disease. It has been ongoing since Florence Nightingale segregated the most battle-injured soldiers during the Crimean War in the 1850s (1). In the 1920s, Dr Walter Dandy organised a three-bed postoperative care unit at Johns Hopkins Hospital, staffed with specialised nurses (1). During World War II, similar specialised sites expanded rapidly when so-called 'shock units' were developed for the postoperative care of the severely injured (2, 3).

In 1953, the Danish anaesthetist Bjørn Aage Ibsen initiated the first intensive care unit (ICU) in Europe after having treated patients during the Copenhagen polio epidemic the previous year (4). During that epidemic, a workforce of about 1500 medical students, nurses and retired nurses treated bulbar poliomyelitis with overpressure ventilation, which reduced mortality rates from above 80% to less than 40% (4-6).

During the following two decades, intensive care medicine was founded and underwent rapid changes. By the end of the 1950s, a four-bed unit called the 'Shock Ward' at the University of Southern California became a prototype for future ICUs. It offered continuous monitoring of the patient's electrocardiogram (ECG), pulse, breathing, central and peripheral temperatures, and arterial and central venous pressures (2).

By 1962 the Shock Ward had a dedicated digital computer system and used algorithms to detect cardiac arrhythmias based on ECG heart rate and pulse rate (2). By today's standard, the system was relatively primitive, but it utilised the available technology in a way we should applaud. Computer technology has been an integral part of intensive care medicine ever since. Intensive care medicine has undergone immense changes since its establishment almost 70 years ago, and advancements in technology have played an essential role in its progress.

## **Modern intensive care medicine**

Today, intensive care medicine is not confined to the ICU as a specific ward. It is a level of care provided prehospital, using well-equipped ambulances and helicopters staffed with anaesthesiologists, and in the emergency department and the general wards before deteriorating patients are transferred to the ICU. The types of patients in the general ICU vary from those with commonly seen conditions, such as sepsis, post-cardiac arrest syndrome (PCAS), respiratory and/or cardiac failure, and trauma, to those with rare diagnoses.

The level of care ranges from observation, based on precautionary principles (e.g. observation for potential airway obstruction), to treating life-threatening multiple organ failure. This care requires close monitoring of each patient 24 hours a day, seven days a week; consequently, the ICU has a higher density of staff and monitoring options compared to a general ward to ensure the safety of these high-risk patients.

The treatment and care are multidisciplinary; several specialities are often involved in the treatment discussion. The staff working in the ICU are specialist nurses and physicians, along with physiotherapists and assistant nurses. In larger hospitals, ICUs are often further subspecialised as paediatric, thoracic or infection ICUs and so forth, while smaller hospitals usually have only one ICU to handle a broad group of patients.

## **Intensive care medicine in Sweden**

Intensive care medicine varies from country to country based on traditions, geographical challenges and healthcare funding. In Sweden, healthcare is mainly tax-funded, and ICU treatment is cost-free for patients. Compared to the rest of Europe, Sweden has a low number of ICU beds per capita, and this number is decreasing. In 2012, Sweden had 5.8 ICU beds per 100,000 inhabitants (7); by 2018, this number had been reduced to 5.1 ICU beds (8). The average in Europe in 2012 was 11.5 ICU beds per 100,000 inhabitants, with Germany having the greatest number of ICU beds per capita with 29.2 per 100,000 inhabitants (9).

The differences in what each country defines as an ICU bed makes direct comparison problematic; nevertheless, a decreasing number of ICU beds in Sweden, when numbers are already at the low end compared to other European countries, is eye-opening and warrants further evaluation. Additionally, because of geographical challenges, access to an ICU bed for the individual patient varies throughout Sweden (9). Healthcare in Sweden is also decentralised, meaning that regional councils are responsible for providing good-quality healthcare based on the central government's guidelines and principles. How this affects the number of ICU beds and care in various regions of Sweden is difficult to say. Differences in prehospital organisations may also affect the overall care of some critically ill patients.

Differences from other Scandinavian countries are notable, with Denmark and Norway having more heavily staffed prehospital setups that include physicians (10). Even within Sweden, there are marked differences in the prehospital organisation (10).

Despite these challenges, the adjusted 30-day mortality for adult patients admitted to an ICU in Sweden has not changed in recent years: the observed mortality rate (OMR) after general ICU admission in Sweden has increased over the last few years, but so has the estimated mortality rate, or EMR (estimation based on the Simplified Acute Physiology Score 3; see below). These two trends balance out each other – patients admitted to a general ICU do not die at greater rates than expected (8).

## ICU prognostication

### **Background**

The ability to predict patient outcome is an important pillar of medicine. Predicting outcome for the general ICU patient is a complicated matter given the wide spectrum of comorbidities, admission circumstances and acute physiological changes on admission. The case-mix of patients admitted to the ICU is very broad, but the common denominator is that they have severe or life-threatening conditions. As intensive care medicine has advanced, several severity scoring models have been developed to predict critically ill patients' outcomes – usually the risk of in-hospital mortality. These scoring systems are not recommended for personalised predictions because of the models' uncertainties at the level of individual patients (11).

Numerous models have been developed to predict mortality after ICU admission (12). The most widely used models are the Acute Physiology and Chronic Health Evaluation (APACHE), the Mortality Probability Model (MPM) and the Simplified Acute Physiology Score (SAPS) (13). They all predict mortality within a specific timeframe but differ in which variables they include and the timeframe in which those variables are obtained. The term often used for the predicted probability of death is the EMR, and the corresponding observed outcome is called the OMR. Other scores, e.g. the Sequential Organ Failure Dysfunction (SOFA) score, are intended to describe the extent of organ dysfunction (14). APACHE, MPM and SAPS have been updated several times since their development almost 40 years ago (13).

The first SAPS model was developed in 1984 based on 679 patients from eight ICUs in France (15). In 1993, SAPS 2 was developed using logistic regression analysis of data from 12,997 patients from 12 countries and 137 ICUs (16). In 2005, SAPS 3 was created based on a complex statistical approach with data from almost 17,000

patients from 303 ICUs across 35 countries (17, 18). APACHE and MPM have undergone similar transitions (13). The latest versions use even more extensive databases than SAPS 3, yet they mainly include ICUs from North America (19, 20). SAPS and MPM remain relatively simple, whereas APACHE has grown increasingly complicated, with more than 100 variables included (19).

SAPS 3 differs from the APACHE score in using data obtained within the time window starting one hour before and ending one hour after ICU admission. By contrast, APACHE uses the worst values recorded for each physiological measure during the first 24 hours after admission. MPM uses both approaches. SAPS 3 further differs from other scoring systems by using a different calibration for predicting hospital mortality in each of seven geographical regions (17, 18). All models above were developed for adult patients admitted to the general ICU, and all predict in-hospital mortality (note that the SAPS 3 model used in Sweden predicts 30-day mortality).

### **ICU prognostication in Sweden**

In Sweden, SAPS 3 is used for ICU prognostication for adults ( $\geq 16$  years of age) admitted to the general ICU. The Paediatric Index of Mortality 3 (PIM3) and a modified Higgins' Intensive Care Unit Admission Score are used for paediatric patients and patients undergoing cardiac surgery, respectively (21, 22).

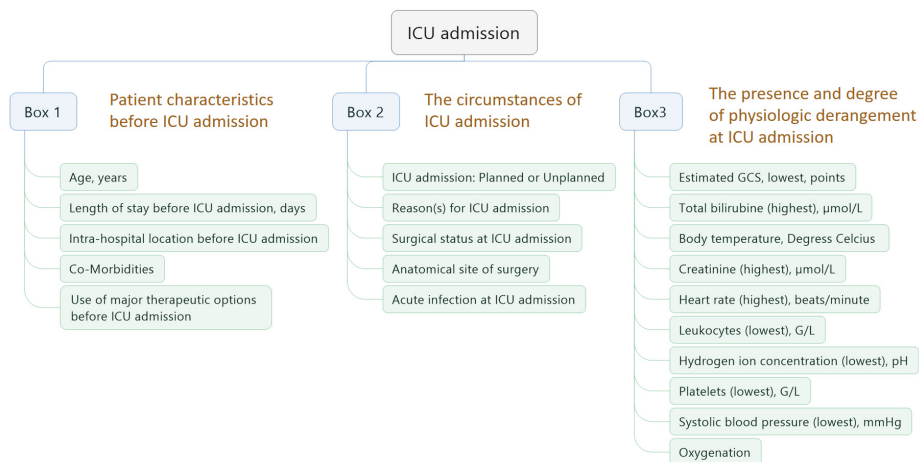
SAPS 3 was introduced in Sweden in 2008, fully incorporated in 2012, and has been calibrated to predict 30-day mortality instead of in-hospital mortality. Furthermore, to assess the status of the patient's central nervous system within SAPS 3, the Glasgow Coma Scale (GCS) was supplemented with the Reaction Level Scale (RLS) to fit medical practice in Sweden (23). SAPS 3 has been calibrated several times since its introduction to optimise its performance for the Swedish ICU population.

The SAPS 3 model consists of three boxes; the total SAPS 3 score is the sum of all boxes (see figure 1):

**Box I:** Patient characteristics before ICU admission: age, comorbidities, location before ICU admission, length of stay in the hospital before ICU admission and the use of major therapeutic options before ICU admission.

**Box II:** Circumstances of ICU admission: reason(s) for ICU admission, anatomic site of surgery (if applicable), whether the ICU admission was planned or unplanned, surgical status and presence of infection at ICU admission.

**Box III:** Presence and degree of physiologic derangement at ICU admission (within one hour before or after admission).



**Figure 1. The third version of the Simplified Acute Physiology Score (SAPS 3).** SAPS 3 consists of three boxes which represent different sets of characteristics. Each variable is transformed into a numeric value, and the sum of these values is the final SAPS 3 score, which can be transformed into a probability of death (the estimated mortality rate).

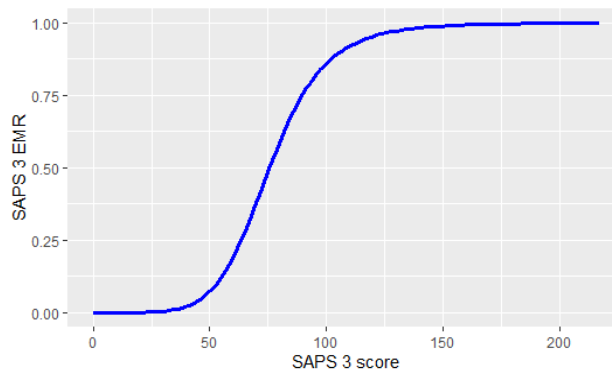
Based on a score sheet, every variable is transformed into a numeric value: age 72 is assigned 13 points, chronic heart failure (New York Heart Association [NYHA] class 4) is assigned 6 points, and so on. The sum of all these values represents the final SAPS 3 score, with a maximum score of 217 points. In the original SAPS 3 model, the final score was then transformed to a probability of death before hospital discharge (in-hospital mortality). In Sweden, however, SAPS 3 has been calibrated to instead predict mortality within 30 days after ICU admission (30-day mortality), which has shown a good discriminative capability, with an area under the receiver operating characteristic curve (AUROC) of 85% (24-26). When SAPS 3 is calibrated to predict 90-day and 180-day mortality, the performance remains good, with AUROCs of 84% and 83%, respectively (26). How to interpret the receiver operating characteristic (ROC) curve is explained within the Methods and materials chapter.

# Improving ICU prognostication

## Recalibration of current models

Calibration of a model is the level of agreement between the model's predictions and the observed outcome and is commonly defined as 'having an event rate of X% among patients with a predicted risk of X%' (27, 28). As changes in the case-mix of ICU patients and in treatment possibilities can affect the calibration over time, it is often necessary to recalibrate a prediction model periodically (29).

Traditionally, a model can be recalibrated by either 1) recalibrating the final score to fit the outcome of choice (level 1 recalibration) or 2) reweighting each variable of the model to fit the outcome of choice (level 2 recalibration, traditionally using logistic regression) (17). The SAPS 3 model has been recalibrated several times (level 1 recalibration) since it was implemented in Sweden. See figure 2 for a visualisation of the latest calibration from 2016 (25). Calibration is further explained within the Methods and materials chapter.



**Figure 2. The latest Swedish SAPS 3 calibration from 2016.** The figure shows how the SAPS 3 score is transformed into a probability of death within 30 days after ICU admission (EMR). The curve characteristics (intercept and slope) can be adjusted during calibration to better fit the current ICU population over time. EMR: estimated mortality rate.

## Adding single predictors

Numerous biomarkers are correlated to ICU mortality, and adding them to SAPS 3 and other scoring systems could potentially improve the predictive performance, overall or for specific diagnoses, of those models. The term 'biomarker' is broadly defined; biomarkers in the following are biochemical biomarkers unless stated otherwise. In this thesis, two biomarkers related to the cardiovascular system are investigated: lactate and high-sensitivity troponin T (hsTnT). Both are affected by the cardiovascular system's load and are easily accessible in the ICU. The levels of these biomarkers are often affected by comorbidities, which can complicate the interpretation.

When a biomarker is added to a prediction model, it is either added to the predicted outcome (EMR) using logistic regression or added to all variables of the original model, thereby creating a new model (often using logistic regression or more recently machine learning methods; see below). The first option is the faster and simpler method and is of an exploratory nature. Studies I and II are both examples of such an approach to adding a biomarker to a prediction model – in this case, SAPS 3. The second option aims to build a completely new model or scoring system from the ground up and is, therefore, more complicated. Studies III–V are examples of such model development (although the objective of these studies was not to investigate the prognostic performance of a specific biomarker). These studies use a machine learning algorithm called an artificial neural network (ANN) and are further explained in the Methods and materials chapter and below.

### *Lactate*

Lactate is a well-known predictor of mortality and is routinely measured in the ICU, as it is included in blood gas analyses. An increased lactate value is a normal occurrence during exercise; a peak value of 15–25 mM can be observed during 'all-out' maximal exercise and lasts 3–8 min post-exercise (30). Similar or lower values observed in the ICU are generally deemed to be highly pathological. The causes of increased lactate vary from disease to disease (31), and levels are affected by co-existing conditions such as liver failure or metformin consumption (32, 33). In patients with seizures, high lactate values are not necessarily related to high mortality, whereas similar values in septic patients would be highly pathological (34). With the latest sepsis guidelines, lactate is now among the criteria for diagnosing septic shock (35), underlining its important role in sepsis. Although lactate values are ubiquitously available, they are not included in SAPS 3.

### *High-sensitivity troponin T*

Cardiac troponin is composed of three subtypes, C, I and T, of which I and T (TnT) are suitable for detecting myocardial injury. It is mainly bound to myofilaments in the cardiomyocytes. The hsTnT assay allows the detection of very low levels of TnT and improves overall diagnostic accuracy in patients with acute myocardial infarction (AMI), which is the main indication for hsTnT (36). In the case of myocardial injury, troponin will begin to rise within three to four hours and remain increased for up to two weeks (37). Increased cardiac troponin is also seen in healthy long-distance runners and is proportional to the distance they run (38, 39). The prognostic value of cardiac troponins for ICU patients seems to vary dependent on the reason for admission (40, 41). Furthermore, it is unclear whether cardiac troponin levels add valuable information when included in today's scoring systems (42–44). Cardiac troponins are not included in SAPS 3.



## Machine learning

Artificial intelligence (AI) has undergone a revolution over the last two decades. Improved algorithms and increased computational power have renewed AI, building on the basic ideas founded in the 1950s. AI is a broad term that can be described as using computers and technology to simulate intelligent behaviour and critical thinking comparable to that of a human being (45). The exact definition is the subject of much discussion and has changed over time due to the rapid developments (46). Two often-used classifications are general AI and narrow AI. General AI refers to machines exhibiting human-like behaviour in their capacity to understand and learn any intellectual task. Narrow AI, on the other hand, focuses on one specific task and is the only type of AI successfully realised today. Narrow AI applications range from autonomous vehicles to image recognition to search engines and much more. In medicine, AI methods have shown promising results in a variety of research fields, from cancer detection to identification of Parkinson's disease to early detection of sepsis (47-49), however, there are still several obstacles before its application can become widespread in various parts of medicine (50, 51).

Machine learning is a branch of AI that focuses on building computer algorithms that improve automatically through experience (52). Machine learning is often categorised as either supervised, unsupervised or reinforcement learning. In unsupervised learning, the algorithm has to find a structure in the input data without knowing the labels (outcome variables). In reinforcement learning, the computer algorithm interacts with an environment while having a specific goal (e.g. playing a game). These types of machine learning are used in critical care research, from investigating fluid treatment strategies in patients admitted with sepsis to discovering subgroups among ICU patients (53-55).

In supervised machine learning (SML), the most widely used machine learning method and the focus of this thesis, the outcome is known to the algorithm (52). SML is designed to learn the relationships and dependencies between input variables (e.g. age or cancer status) and an output variable such as 30-day mortality (classification) or length of stay (regression).

Methods such as support vector machines, the naive Bayes algorithm, the k-nearest neighbours algorithm, decision trees and ANNs are all SML algorithms (56, 57). Even traditional statistical methods used in medicine – logistic regression and linear regression – can be classified as SML algorithms (57, 58). SML methods generally perform well, but performance may depend on the type of data they are interpreting (59). When the input variables have complex interactions between them, certain algorithms such as ANNs may work better based on their design. The ANN is one out of many SML methods, and sometimes ANNs are combined with other SML methods to improve prognostic performance (ensemble learning). ANNs are used in studies III–V and are briefly introduced in the Methods and materials chapter.

In recent years, numerous studies have investigated whether the various types of machine learning can improve mortality and morbidity predictions in the ICU. Studies have reported promising results in predicting length of ICU stay, instability in the ICU, risk of developing kidney injury, pulmonary emboli, and so forth using various SML methods (60). Deasy et al. showed how a dynamic model could improve mortality prediction over time after ICU admission (61). Similar findings were reported in a recent study by Thorsen-Meyer et al. (62). The designs of both studies differed from the approach used by the SAPS 3 model, as they used data obtained after ICU admission. To our knowledge, no published studies have used ANNs to interpret admission data based on the SAPS 3 variables and compared its performance to that of the SAPS 3 model in predicting 30-day mortality (which was the aim of study III).

### **Diagnosis-specific predictions**

While overall scoring systems such as SAPS 3 have to perform well for general ICU populations, some ICU diagnoses might benefit from being managed independently. Studies IV and V focus on long-term neurological prognostication for comatose patients admitted to the ICU after out-of-hospital cardiac arrest. Similar diagnosis-specific approaches could be used for other patient groups as well, such as patients with trauma, sepsis or respiratory failure.

## **Post-cardiac arrest syndrome and prognostication**

Patients admitted post-cardiac arrest constitute a distinct group in the ICU and an epitome of when prognostication has direct consequences. When prognostication is necessary, these patients are still comatose and ventilated. Based on the prognostication, one of two things can happen: either treatment will continue, or a decision will be made to withdraw life-sustaining therapy (WLST). This means very accurate and secure prediction models are needed to avoid false-positive predictions (predicted poor outcome, reported good outcome).

Cardiac arrest is the abrupt loss of heart function, which may be reversible or may lead to death. Cardiac arrest can be categorised as in-hospital or out-of-hospital cardiac arrest (OHCA). This thesis focuses on OHCA. In Europe and the United States, the incidence of OHCA with resuscitation attempted is between 50 and 110 per 100,000 person-years (63, 64). Overall survival to hospital discharge is 10%–12% in Europe and the United States, with wide variation among individual countries (63, 64). In Sweden, 30-day survival is approximately 11%, a rate which has more than doubled since 2000 (65).

The duration of resuscitation is an important determinant of survival and can be divided into two intervals: 1) no flow (from cardiac arrest to initiation of cardiopulmonary resuscitation [CPR]) and 2) low flow (from the start of CPR to return of spontaneous circulation [ROSC] or the termination of resuscitation) (66). It is fundamental to begin CPR as soon as possible and defibrillate if the rhythm is shockable to increase the possibility of a good outcome (67). Furthermore, the survival rate is better in patients with a shockable rhythm (65).

### **Post–cardiac arrest syndrome**

The injuries from the periods of no flow (NF) and low flow (LF) are often not directly reversed by ROSC. The reperfusion may even cause damage on its own. The term PCAS is used to describe these complex processes; PCAS has four components (68):

- 1) post–cardiac arrest brain injury,
- 2) post–cardiac arrest myocardial dysfunction,
- 3) systemic ischemia/reperfusion response and
- 4) persistent precipitating pathology.

The patient can wake up directly if the time to ROSC is very short (defibrillation within a few minutes). However, the general management of PCAS requires ICU admission for general ICU treatment and monitoring, including advanced cerebral and haemodynamic monitoring (68). The initial objectives are to initiate targeted temperature management (TTM), optimise mechanical ventilation and haemodynamics, and identify and treat acute coronary syndrome along with other causes of cardiac arrest, in addition to providing standard intensive care (69).

### **Post–cardiac arrest prognostication**

Neurological outcome after cardiac arrest varies from no symptoms to a vegetative state. Both the Cerebral Performance Category scale (CPC; see table 1) (70, 71) and the modified Rankin Scale (mRS) (72) are frequently used to classify the neurological outcome. Both are often, furthermore, dichotomised into poor outcome or good outcome. This thesis focuses primarily on the CPC.

**Table 1. Cerebral Performance Category scale (CPC).**

Dichotomised outcome in this thesis	CPC
Good outcome	<b>CPC 1: Good cerebral performance:</b> Conscious, alert, able to work, might have a mild neurologic or psychologic deficit.
	<b>CPC 2: Moderate cerebral disability:</b> Conscious, sufficient cerebral function for independent activities of daily life. Able to work in a sheltered environment.
Poor outcome	<b>CPC 3: Severe cerebral disability:</b> Conscious, dependent on others for daily support because of impaired brain function. Ranges from ambulatory state to severe dementia or paralysis.
	<b>CPC 4: Coma or vegetative state:</b> Any degree of coma without the presence of all brain death criteria. Unawareness, even if the patient appears awake (vegetative state) without interaction with the environment; may have spontaneous eye-opening and sleep/wake cycles. Cerebral unresponsiveness.
	<b>CPC 5: Brain death:</b> Certified brain dead or dead by traditional criteria.

Models such as the TTM risk score and the Miracle<sub>2</sub> score aim to predict the six-month neurological outcome at hospital admission following OHCA (73, 74). Both models base their predictions on information obtained prehospital and on hospital admission, and have reported good prognostic performance with AUROCs of 84% and 88%, respectively (74, 75). The TTM risk score is used for comparison in study IV.

Among OHCA patients who die shortly after ICU admission, the main cause of death is cardiac failure, whereas neurological injury accounts for the majority of later deaths (76). Among OHCA patients, two-thirds of all deaths before hospital discharge are due to neurological injury (77). Most of these deaths occur after WLST as a consequence of prognostication. As mentioned above, this demands accurate prognostication, primarily to avoid false-positive predictions (predicted poor outcome, reported good outcome) and secondarily to ensure a low number of false-negative predictions (predicted good outcome, reported poor outcome).

It is recommended that the prognostication of comatose post-cardiac arrest patients be delayed for at least 72 h post-cardiac arrest and that it be multimodal according to the current guidelines (78). The clinical neurological examination, neurological imaging, electroencephalogram (EEG), short-latency somatosensory evoked potentials (SSEP) and biomarkers are important methods for evaluating the extent of brain injury. The theory and evidence behind these techniques are complex and are mentioned here only briefly, as these are not the focus of this thesis.

- Neurological imaging such as computed tomography (CT) or magnetic resonance imaging (MRI) can be a valuable tool. CT is often performed within 24 h after cardiac arrest to exclude haemorrhages or other pathologies. MRI is recommended two to five days after ROSC and is used to detect whether the patient has cerebral oedema.
- EEG is frequently used in post-cardiac arrest prognostication. It can be performed continuously (cEEG), producing less information, or as an intermittent full examination. In the current ERC–ESICM guidelines (see below), two EEG patterns are related to poor outcome (78): unreactive burst-suppression (EEG patterns with intermittent periods of low-voltage electric activity for more than 50% of the EEG alternating with irregularly high-voltage electric activity) and unreactive status epilepticus (78, 79).
- When using SSEP, the N20 potential is the contralateral response in the primary somatosensory cortex to stimuli of the median nerve. When absent bilaterally, this is considered a robust predictor of poor outcome (80).
- Biomarkers are described later in the chapter.

The latest prognostication guidelines from the European Resuscitation Council and European Society of Intensive Care Medicine (ERC–ESICM), from 2015, recommend a four-step model used at least 72 h after ROSC (78). It focuses only on patients with no motor response or extension from pain (Glasgow Coma Scale motor response score [GCS-M]  $\leq 2$ ) after confounders are excluded (particularly residual sedation). If the patient has no pupillary and corneal reflexes or if the N20 potential on the SSEP is absent (bilateral examinations), a poor outcome is very likely. Otherwise, the next step is to re-evaluate the patient after 24 h. If the patient then still has a GCS-M  $\leq 2$ , the prognostication should continue, and if two or more of the following are present, a poor outcome is likely: status myoclonus  $\leq 48$  h after ROSC, high neuron-specific enolase (NSE) values, unreactive burst-suppression or status epilepticus on EEG, and diffuse anoxic injury on brain CT or MRI. If none or only one of these criteria is present, then the patient should be observed and evaluated later.

Three studies have recently assessed the accuracy of the ERC–ESICM guidelines, and all three reported a 0% false-positive rate (FPR) (81-83). The consistency is reassuring and naturally leads to these questions: how to decrease the false-negative rate (FNR) and how to broaden the inclusion criteria while retaining a 0% FPR (84).

## **Biomarkers in cardiac arrest prognostication**

Numerous biomarkers have been investigated for use in improving post-cardiac prognostication. These biomarkers focus on one of the following components of PCAS:

- Brain injury: NSE, neurofilament light (NFL), S100 calcium-binding protein B (S100B), tau protein (tau), glial fibrillary acidic protein (GFAP) and ubiquitin carboxy-terminal hydrolase L1 (UCHL1) (85-91).
- Cardiac injury: TnT, N-terminal pro-B-type natriuretic peptide (BNP) and copeptin (92, 93).
- Systemic inflammation: Procalcitonin (PCT) and interleukin-6 (IL-6) (94, 95).

This list of biomarkers represents the biomarkers used in this thesis. Their predictive abilities vary, and little is known about their prognostic values when combined (96). As a single biomarker, NFL is the most promising, with an AUROC of 94%–98% when predicting a poor neurological outcome as early as 24 h post-OHCA (97, 98). NFL is a protein highly expressed in large-calibre myelinated axons and has shown promising results in detecting the degree of axonal damage in various neurological disorders (97).

NSE is a highly specific marker for neurons and peripheral neuroendocrine cells and can be used in cancer diagnosis (neuroendocrine tumours) and in estimating the extent of brain injury (99). It is important to acknowledge that haemolysis can result in false-high NSE values due to the high NSE concentration in erythrocytes (100). NSE is not reliable as a prognostic biomarker 24 h post-cardiac arrest, but its accuracy improves after 48 h and 72 h (101, 102). NSE is the only biomarker recommended in the current guidelines (78).



# Aims of the thesis

- I. To investigate the prognostic value of lactate obtained on admission when combined with SAPS 3 (Swedish modification) for ICU patients overall and for patients with the five most common primary ICU diagnoses.
- II. To investigate whether hsTnT obtained on ICU admission improves the prognostic accuracy of SAPS 3 (Swedish modification) for 1) ICU patients in general, 2) cardiac arrest patients and 3) non-cardiac arrest patients, particularly patients with the three most common diagnoses in this group.
- III. To improve 30-day mortality prognostication by using ANNs to interpret the variables used in the SAPS 3 model (Swedish modification) and to identify the smallest possible subset of SAPS 3 variables which can retain the same performance as the full SAPS 3 model.
- IV. To use ANNs to create a model for early prediction of long-term neurological outcome for comatose survivors of OHCA admitted to the ICU, and to use this model to investigate the intervention effect in cardiac arrest patients treated with TTM.
- V. To investigate whether cumulative information obtained during the first three days of intensive care can, when processed with ANNs, produce a reliable model for predicting neurological outcome post-OHCA with and without clinically accessible and research-grade biomarkers.





# Methods and materials

The first and second papers incorporated into this thesis are based on data from the general ICU at Skåne University Hospital in Lund, Sweden, and PASIVA, a patient administration system which collects data from ICUs. The third paper is based on national data from the Swedish Intensive Care Registry (SIR) only, while the fourth and fifth papers are post hoc analyses of the Target Temperature Management trial (TTM-trial) (103). This chapter summarises the materials and methods described in papers I–V. Detailed descriptions of the methods and materials used in the five papers are presented in the respective papers. These methods and materials are summarised in table 2, which also includes the number of participants for the final analysis in each study.

**Table 2. Overview of the five studies included in the thesis.**

Intensive care unit (ICU), Cerebral Performance Category scale (CPC), estimated mortality rate (EMR), high-sensitivity troponin T (hsTnT), Simplified Acute Physiology Score 3 (SAPS 3). \*Participants for final analysis. \*\*The number of participants in study V varied based on the time point and biomarkers of interest.

Paper	I	II	III	IV & V
Design	A single centre retrospective study	A single centre retrospective study	A national multicentre retrospective study	Post hoc analysis of an international randomised multicentre trial
Study population	General ICU population	General ICU population	General ICU population	Specific ICU population: Comatose survivors of out-of-hospital cardiac arrest from a presumed cardiac cause
	Adults	Adults	Adults	Adults
	January 2008 – June 2017	February 2010 – June 2017	2009–2017	2010–2013
Participants*	n = 3039	n = 856	n = 217,289	n = 932**
Variables	SAPS 3 EMR + lactate on ICU admission	SAPS 3 EMR + hsTnT on ICU admission	SAPS 3 variables	<u>Study IV</u> : Background, prehospital and admission variables. <u>Study V</u> : Similar to study IV and at days 1, 2 and 3 after ICU admission.
Outcome	30-day mortality	30-day mortality	30-day mortality	Binary 6-month neurological outcome: CPC 1–2 or CPC 3–5
Method	Logistic regression	Logistic regression	Artificial neural networks	Artificial neural networks

## Sources of data

### PASIVA

PASIVA (*Patientadministrativt system för Intensivvårdsavdelningar*) is a patient administration system which collects data from 61 out of 84 ICUs in Sweden. Originally, PASIVA was designed to forward data to SIR, but it later evolved to also provide feedback to end users (the ICUs) as a tool to improve and plan the care of patients.

### Swedish Intensive Care Registry

SIR is a non-profit organisation which prospectively collects patient-level data for all patients admitted to an ICU in Sweden. The members of SIR are the Swedish ICUs and represent a wide array of ICUs: general, neurosurgical, thoracic surgical, burn, infection and paediatric ICUs, and one extracorporeal membrane oxygenation (ECMO) centre. Since SIR was established in 2001, it has gradually grown in size; in 2020 all of the 84 ICUs in Sweden were members of SIR. Sweden has 526 ICU beds (5.1 ICU beds per 100,000 citizens) and has approximately 45,000 ICU admissions every year. The main purpose of SIR is to utilise data to improve the care of ICU patients (8).

The data we used from SIR contained basic information about each admission and patient, SAPS 3 input, survival data (originally from the Swedish National Population Register), and a primary diagnosis for each admission, which was based on a subset of the Swedish version of the 10<sup>th</sup> revision of the International Classification of Diseases (ICD-10). Note that this subset of ICD-10 diagnoses is no longer being used for primary diagnoses.

### The Target Temperature Management trial

The TTM-trial was an international randomised multicentre trial designed to find a difference in survival among comatose cardiac arrest survivors treated with different target temperatures after ICU admission. Patients were enrolled from November 2010 to January 2013 from 36 ICUs across Europe and Australia (103). The fourth and fifth papers are post hoc analyses of the TTM-trial.

#### *Trial design*

The inclusion criteria allowed participation of comatose ( $GCS \leq 8$ ) adults ( $\geq 18$  years of age) with a sustained ROSC after resuscitation from OHCA of a presumed cardiac cause.

Exclusion criteria were

- limitations in therapy, including do-not-resuscitate orders, or known illness making survival to 180 days unlikely
- suspected or known acute intracranial haemorrhage or stroke
- pre-existing neurological disability (CPC 3–4)
- unwitnessed cardiac arrest with asystole as the initial rhythm
- persistent cardiogenic shock despite medical interventions and mechanical assist
- a body temperature of less than 30°C
- previous bleeding diathesis
- pregnancy
- more than 240 minutes from ROSC to screening (104).

Patients were randomised to a target temperature of either 33°C or 36°C for a total of 28 h of temperature management using invasive or surface cooling, followed by gradual warming to 37°C at 0.5°C/h and avoidance of a body temperature above 37.5°C until 72 h after OHCA. According to the study protocol, the neurological prognostication was performed at least 108 h post-cardiac arrest (72 h after rewarming).

The primary outcome was all-cause mortality until the end of the trial, and the secondary outcome was the neurological outcome (including death) at six months measured by the mRS and the CPC scale (assessor-blinded) (104).

### *Results*

No difference was found in the two treatment arms in terms of all-cause mortality (primary outcome) or neurological outcome (secondary outcome) (103).

# Methods

The methods for all studies are briefly described in this section. The concepts of training, validation and test datasets; ANNs; and performance measures used in this thesis are introduced briefly here to provide a better understanding.

## Performance measures

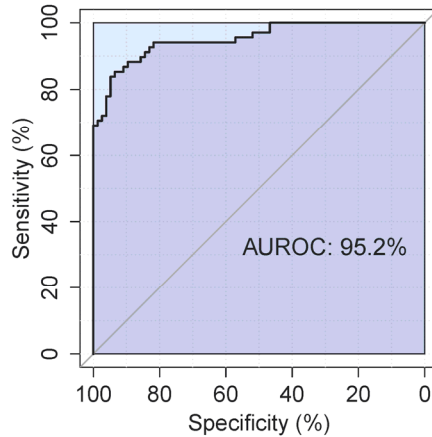
Because in-hospital mortality and 30-day mortality are indicated with yes or no answers, where survival can be classified as 0 and death as 1, predicting this type of outcome is called binary classification. A prediction model does not give a yes (deceased) or no (survived) prediction but provides a probability between 0 (survived) and 1 (deceased): for example, a patient may have a 30% probability of dying within the next 30 days. As mentioned earlier, this prediction is called the EMR.

There are numerous performance measures for binary classification, some of which are introduced here. When examining the performance of these predictive models, both discrimination and calibration are essential.

### *Discrimination*

Discrimination describes the ability of the model to distinguish between two outcomes – for example, survivor and non-survivor – and is measured using the AUROC as exemplified in figure 3. The AUROC has a value between 0 and 1; the closer the AUROC is to 1, the better the model can classify whether a patient will survive or die. If the AUROC is 50%, the prediction is no better than a completely random selection. An AUROC of 70%–80% can be classified as fair, 80%–90% as good and 90%–100% as excellent, even though the ranges of these labels vary (105). Moreover, when evaluating discrimination, it is important to look at the layout of the ROC curve as well to inspect the specificity and sensitivity of the model (see below). Even if the AUROC is high, the model might not work as intended.

The ROC curve describes the model's classification capability at different thresholds; the true-positive rate (TPR) and the false-positive rate (FPR) are calculated at various thresholds and plotted as in figure 3. The TPR is also called sensitivity (and recall) and is plotted on the  $y$ -axis. The  $x$ -axis in the figure displays  $1 - \text{FPR}$ , also called specificity. The ROC curve provides additional information beyond the AUROC, as it describes the trade-off between specificity and sensitivity. In the example in figure 3, the sensitivity is approximately 70% when the specificity is at 100%.



**Figure 3. The receiver operating characteristic (ROC) curve.** The sensitivity (%) is plotted on the y-axis, and the specificity (1 - false-positive rate) is plotted on the x-axis. The area under the receiver operating characteristic curve (AUROC) in this example is 95.2%.

Both FPR and TPR are calculated based on the confusion matrix for the chosen threshold. The confusion matrix for a binary classification problem is a 2x2 table which displays the performance of a prediction model at a specific threshold between 0 and 1.

**Table 3. The confusion matrix.** The rows represent the prediction, and the columns represent the observed outcome.

	Observed Positive	Observed Negative
Predicted Positive	True positives (TP)	False positives (FP)
Predicted Negative	False negatives (FN)	True negatives (TN)

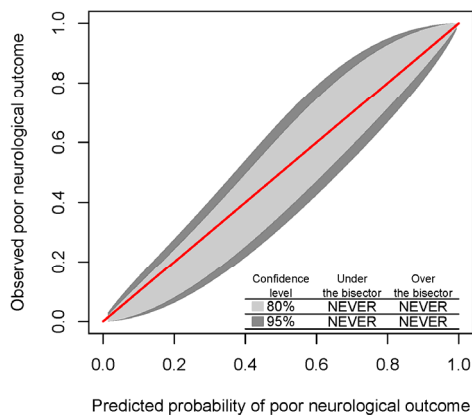
As seen in table 3, each prediction can be either a true positive (TP), a false positive (FP), a false negative (FN) or a true negative (TN). These four measures are the foundation for calculating numerous other performance measures for binary classification which are not used in this thesis.

### *Calibration*

As mentioned previously, calibration is the level of agreement between the predictions and the observed outcome. A calibration plot is often used to describe the calibration, with the prediction on the x-axis and the outcome on the y-axis. As the outcome is binary, the predictions are usually plotted by decile on the x-axis, with the corresponding observations on the y-axis (106). The diagonal of a calibration plot represents the perfect calibration, where the prediction on the x-axis

correlates perfectly with the corresponding observation. Finazzi et al. developed this measure even further, adding a calibration belt with a confidence interval (CI; e.g. 95%), as seen in figure 4 (107). This GiViTI (Italian Group for the Evaluation of Interventions in Intensive Care Medicine) calibration belt also reveals when the model is under- or overestimating the risk (under or over the bisector). Data from a sufficient number of patients is required to create a precise calibration curve; a minimum of 200 patients with and without the outcome has been suggested (28). The example in figure 4 uses only 145 patients (study V), which is why it was not included in the original paper and why the CIs are wide.

Standardised mortality ratio (SMR) is an often-used overall measure, defined as the ratio between the observed and the expected numbers of deaths (OMR/EMR). As a result, if the SMR is equal to 1, the observed number of deaths is as expected; if the SMR is greater than 1, the mortality is higher than expected and vice versa (108). Other measures of calibration, such as the Cox calibration test and the Hosmer–Lemeshow test, are not used in this thesis.



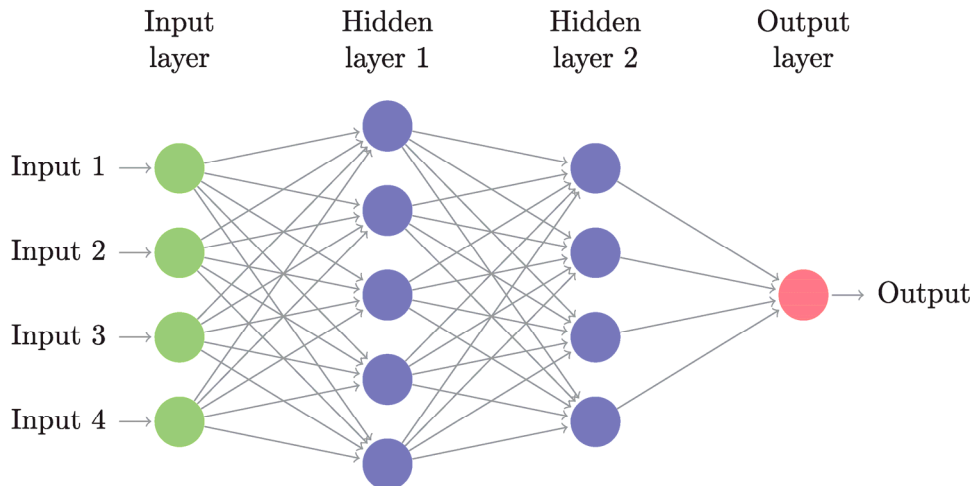
**Figure 4. Calibration plot.** An example of a calibration plot (from study V) with the GiViTI (Italian Group for the Evaluation of Interventions in Intensive Care Medicine) calibration belt.

### *Others measures*

The Brier score is used to describe the overall model performance. It is a calculation of the squared difference between the predicted probabilities and the actual outcomes. The best possible score is 0, the worst is 1 (totally inaccurate), and a score of 0.25 can be expected to occur by chance (109).

## Artificial neural network (ANN) – A brief introduction

The theory behind ANN is complex. The basic idea is to mimic the structure of the neurons in the human brain. As in the human nervous system, the nodes (the neurons of an ANN) are linked together and can receive, transform and send information forward. An ANN consists of an input layer, a number of hidden layers and an output layer, as illustrated in figure 5.



**Figure 5. Artificial neural network (ANN).** A schematic ANN with an input layer with four nodes, two hidden layers with five and four nodes and an output layer with one node. All nodes are connected to the previous and next layers by weights. In total, an ANN as seen above will have 54 weights when the bias nodes (not shown) are included.

The number of nodes in the input layer is equal to the number of variables chosen for the model. The numbers of nodes in the hidden layers can vary, and the output layer consists of one node (when the ANN is performing binary classification).

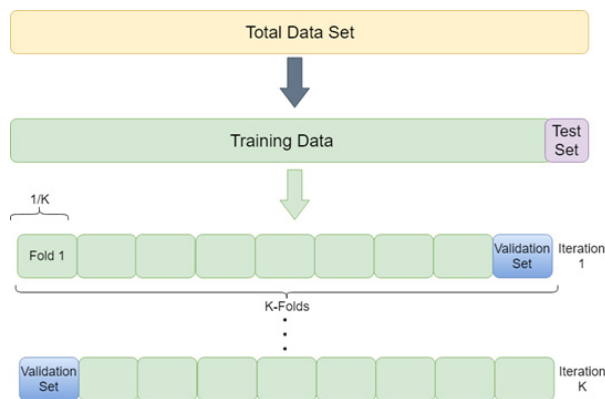
Weights link each node to all nodes in the previous layer and the next layer. As the model is trained to learn the relationships and patterns between the input and output variables, these weights are adjusted to optimise the prediction. The goal is to minimise the difference between the ANN prediction and the actual outcome during training, also called error.

Two concepts, ‘gradient descent’ and ‘backpropagation’, are important to be able to understand how the weights are adjusted. Gradient descent is an optimisation algorithm in which the error is gradually minimised. To minimize the error, the weights throughout the ANN are simply increased or decreased based on the gradient of each individual weight. Backpropagation is a method that allows for an efficient calculation of the gradients, starting with the output layer and then moving backwards.



It is important to understand that the weights are adjusted only by the algorithm; this adjustment is the way the computer learns the patterns between the input and output variables. The parameters controlled by the user are called hyperparameters and include the number of hidden layers, the number of nodes in each layer, how each node handles the information from the previous layer (activation function) and how many times the dataset passes through the network (number of epochs) and more. Selecting the optimal hyperparameters is essential to ensuring the best possible conditions to achieve the best prediction model. The search to find the best possible value for a hyperparameter can be done manually or can be automated using a grid search, random search or Bayesian optimisation.

An ANN differs from logistic regression by being highly adaptable when finding patterns in the data. This property of ANNs is both a strength and a weakness, as it can lead to overfitting. When a model is overfitted to the training data used for model development, it exhibits poor generalisability when tested on other patient populations. There are several tools to avoid overfitting; a fundamental one is splitting the dataset into a test set and training set.



**Figure 6. Training, validation and test sets.** The figure illustrates how the data is split into a training set and a test set, and how the training set can be split into different validation sets (*k*-fold cross-validation).

The dataset is divided into a training set for model development and a test set for use in an unbiased evaluation of the final model (see figure 6). Furthermore, a part of the training set can be allocated as a validation set. A validation set is a sample of the training set held back when training the model; it is then used to give an estimate of model performance while determining the best possible hyperparameters (also called hyperparameter tuning). The validation set may be used to stop training (after a number of epochs) so that overfitting does not occur. In this way, the validation set is a part of the model development and can, therefore, not be used for an unbiased evaluation. The test set, by contrast, is not used during model development and is set aside for an unbiased performance evaluation of the final model.

Changing the validation set during training based on predefined splits – so-called *k*-fold cross-validation – is commonly employed (see figure 6). Before the final model is implemented as a prediction model, it should be validated on an entirely new set of patients (external validation), as the test dataset is associated with the data used for model development.

The complexity of ANNs (i.e. their large numbers of parameters) makes interpretation difficult. Although all information about the structure and the weights of the ANN can be extracted, it is far too complicated for the human mind to comprehend. This complexity has raised concerns about using black-box models. However, new methods such as Shapley additive explanations (SHAP) can help explain the output of various SML algorithms, including ANNs (110, 111).

### **Value of additional biomarkers (studies I and II)**

In these studies, we investigated the prognostic performance of SAPS 3 when biomarkers were added. Both studies used the Swedish SAPS 3 calibration from 2016 to calculate the EMR:

$$EMR = \frac{e^{Logit}}{(1 + e^{Logit})},$$

where

$$Logit = -32.06302 + \ln(SAPS\ 3\ score + 10.34171) * 7.199704.$$

See the Background chapter for a visualisation of the EMR calculation shown above.

#### *Study I*

In this retrospective study, we investigated all adult admissions to the general ICU at Skåne University Hospital in Lund, Sweden, between 1 January 2008 and 30 June 2017. The highest lactate value within one hour of ICU admission was obtained from medical records. Our ICU used a regularly calibrated Radiometer ABL 800 Flex blood gas machine (Radiometer, Copenhagen, Denmark) to measure lactate concentrations.

Statistical analysis: We included the lactate value and the SAPS 3 EMR as variables in a multivariate logistic regression on 30-day mortality. To quantify the discrimination capability of SAPS 3 when lactate was added, we calculated the AUROC for diagnoses with significant odds ratios. The AUROCs were compared using DeLong's test (112).

## *Study II*

In this retrospective study, we investigated all adult admissions to the general ICU at Skåne University Hospital in Lund, Sweden, between 25 February and 30 June 2017. The highest hsTnT value within 1.5 h of ICU admission was obtained from medical records. A Cobas 8000 analyser (Roche, Germany) was used to measure hsTnT.

Statistical analysis: We included hsTnT and SAPS 3 EMR as variables in a multivariate logistic regression on 30-day mortality. To quantify changes in discrimination caused by adding hsTnT to SAPS 3, we calculated the AUROCs for the following groups of diagnoses: overall, cardiac arrest and non-cardiac arrest (sepsis, heart failure and respiratory failure). DeLong's test was used to compare the AUROCs.

## **Using ANNs to short-term predict mortality (study III)**

From SIR we identified all first-time adult ICU admissions with at least 30 days of follow-up data during the period of 2009–2017. Cardiothoracic admissions were excluded, as they use a different scoring system. All variables used to calculate SAPS 3 and 30-day mortality were used in this study.

**Model development:** One-sixth of the dataset was selected at random and set aside for independent validation purposes (the test set), and the rest of the dataset (5/6) was used for model development. We used a grid search of 200 ANNs using two hidden layers and varying numbers of nodes (between 5 and 400) for each ANN to find the best possible structure to predict 30-day mortality. Batch normalisation was used to improve training speed and accuracy. The loss function was optimised using the Adam implementation of stochastic gradient descent, using a learning rate of 0.001. To increase generalisability, we used drop-out in the input and the hidden layers, and used five-fold cross-validation during model development. We used 100 epochs during training for each network with a batch size of 512 using the rectified linear unit (ReLU) function in the hidden layers. We used mean and mode to impute missing values and measured model performance using the AUROC, calibration curve and Brier score. DeLong's test was used to compare the AUROCs of the two models. As a final step, we used ANNs to identify the smallest possible subset of SAPS 3 variables which could retain the same level of AUROC performance as the full SAPS 3 model.

## **Predicting neurological outcome after out-of-hospital cardiac arrest (studies IV and V)**

For studies IV and V, we included all patients from the primary analysis of the TTM-trial (n = 939) (103). In both studies, we excluded patients with missing six-month neurological outcomes and patients with a large number of missing values (>40 missing values at hospital admission). Hence, the initial patient population in both studies was the same. In study V, the patient population subsequently changed depending on the time point after ICU admission; at time points 24 h, 48 h and 72 h, patients who had awoken or died were excluded. The outcome in both studies was the neurological outcome at six months, including survival, using a dichotomised CPC scale, with CPC 1–2 categorised as a good functional outcome and CPC 3–5 as a poor functional outcome.

### *Study IV*

We created a prediction model for comatose OHCA patients that was based on information available at ICU admission: background, prehospital and admission data.

**Model development:** We randomly set aside 10% of the dataset to test the performance of the final model (the test set) and used the remaining data (90%) for model development. We used drop-out and five-fold cross-validation during model development, and we used a Bayesian optimisation approach to find the best network structure (hyperparameters). See table 4 for the limits used during the search for the best hyperparameters. All networks were trained with early stopping with a patience of 50 epochs. The final model was chosen based on the AUROC of the cross-validations. The AUROC was reported using the test set data and was then compared to a logistic regression-based model's AUROC (after removing patients who originally had missing values) using DeLong's test. The final model was also used to investigate the effect of TTM at 33°C vs 36°C based on patients' risk stratification.

In the search for a simplified model, we ranked all input variables by subtracting one variable at a time from the developed model and calculating the AUROC. We then started with the most important variables from this ranking and added one variable at a time back to the model, recalculating the AUROC based on the training set at each step.

**Table 4. Hyperparameters during model development in study IV.** The predefined limits for hyperparameter tuning during development to find the best possible model using Bayesian optimisation.

Hyperparameters	Limits during model development
Number of hidden layers	1–4
Nodes in each layer	5–400
Batch size	1–128
Drop-out rate	0–0.3 for the input layer and 0–0.5 for the hidden layers
Norm regularisation	L <sub>1</sub> , L <sub>2</sub> or Max-norm
Activation function for the hidden layers	Rectified linear unit (ReLU) or hyperbolic tangent function
Optimisation	Adam implementation of stochastic gradient descent or a slightly different version called Adam AMSgrad

Finally, to investigate whether the patient's risk group would determine whether one of the two target temperatures would be beneficial, the cohort was divided into five classes of poor outcome risk.

### *Study V*

This study was a sequel to study IV that used cumulative clinical variables along with clinically accessible and research-grade biomarkers gathered during the first three days after ICU admission. We used biomarkers from the TTM-trial biobank, which had collected blood samples from 29 of the 36 trial sites, and categorised them into additional levels of biomarkers beyond the level of the clinical variables already available from the TTM-trial database:

- Level A: Clinical variables only
- Level B: Level A plus clinically accessible biomarkers: NSE, S100B, TnT, BNP and PCT
- Level C: Level B plus research-grade biomarkers: NFL, copeptin, IL-6, tau, GFAP and UCHL1.

In total, nine datasets were created: three levels of biomarkers each from 24 h (day 1), 48 h (day 2) and 72 h (day 3) after ICU admission.

The datasets were randomly divided into a training set for model development (80%) and a test set for internal validation (20%). The randomisation key was created at the time of hospital admission; hence the split was the same for all models. The number of variables were reduced in each dataset by using a correlation threshold of 98%, a missing values threshold of 20%, a minimum incidence of 2% for unique binary variable events, and a wrapper variable selection method which combined the feature selection algorithm with Shapley values. Missing values were imputed using median and mode imputation for continuous and binary variables, respectively. More advanced imputation methods such as ‘missforest’ did not outperform median/mode imputation (data not shown).

**Model development:** An ANN model was developed for each of the nine datasets. Similar to study IV, we again used five-fold cross-validation and a Bayesian optimisation algorithm to find the most suitable hyperparameter values (see table 5). Early stopping was applied with a patience of 30 epochs to avoid overfitting, and we used a fixed learning rate of  $10^{-3}$ . Model performance was reported by calculating the AUROC for the test set data and displaying the ROC curve for all models.

To find the optimal probability threshold for cardiac arrest prognostication, we based the threshold on 100% specificity in the training set. We then reported the distribution of the confusion matrix based on the test set.

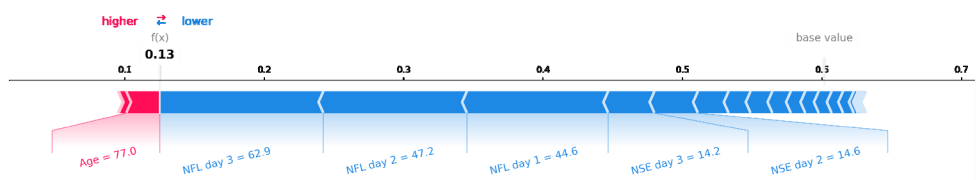
**Table 5. Hyperparameters during model development in study V.** The predefined limits for hyperparameter tuning during development to find the best possible model using Bayesian optimisation.

Hyperparameters	Limits during model development
Number of hidden layers	1–3
Nodes in each layer	5–250
Batch size	4–128
Drop-out rate	0–0.5 for the input layer and 0–0.5 for the hidden layers
Norm regularisation	L <sub>1</sub> , L <sub>2</sub> or Max-norm
Activation function for the hidden layers	ReLU or hyperbolic tangent function

### Shapley additive explanations algorithm

The complexity of ANNs and other advanced machine learning algorithms can be a barrier to implementation in a clinical setting. In study V, we applied the SHAP algorithm to explain how the individual predictions were attained (110, 111). The SHAP algorithm is based on Shapley values, which originate from game theory.

The basic idea is to explain how much a single variable contributes to the final prediction based on that variable’s effect on the difference between the actual prediction and the mean of all predictions. Each prediction starts at the mean value of all predictions (the baseline). Each variable either increases or decreases the risk. These SHAP values, also called ‘forces’, balance each other out in the actual prediction (113).



**Figure 7. SHAP explanation force plot.** The patient in this example was predicted to have a 13% risk of a poor outcome. The patient’s age was the factor that contributed most to increasing the risk (marked with red), and modest levels of NFL and NSE contributed most to decreasing the risk (marked with blue). NFL: neurofilament light (ng/L); NSE: neuron-specific enolase (ng/mL); SHAP: Shapley additive explanations algorithm.

Figure 7 is an example of a ‘SHAP explanation force plot’. The base value, the mean of all predictions, is approximately 0.5, and the patient’s predicted risk of a poor outcome (in this case, poor neurological outcome after six months) is 0.13. Biomarkers such as NFL and NSE (marked with blue) decrease this particular patient’s risk of poor outcome, and the patient’s age is the most important factor that increase that risk (marked red).

The SHAP algorithm can also help to explain how the entire prediction model works. By calculating Shapley values for all patients (and thereby creating a matrix of Shapley values), we can further interpret the prediction model. In study V, we used the ‘SHAP feature importance’ to rank the most important variables as measured by the mean absolute Shapley values.

## Software

The statistical analysis was performed using R, version 3.2.3–4.0.0 (R Foundation for Statistical Computing), and Python, version 3.6.4–3.8.3 (Python Software Foundation) (114, 115). All ANN models in the thesis were developed using Tensorflow, an open-source framework developed by Google (116).

The ‘tableone’ package was used to calculate the differences in the study populations (117). The ‘forestplot’ package in R was used to display the odds ratio (118). ROC curves and AUROC calculations were performed using the ‘pROC’ package in R (119). The ‘Optimalcutpoints’ package in R was used for calculating thresholds for the confusion matrix in study V (120). The ‘Boruta–Shap’ and ‘shap’ packages in Python were used for variable selection and explanation of the ANN model in study V, respectively (110, 121). The schematic ANN figure was created using the ‘TikZ’ package (122).

## Ethics

Studies I–III were approved by the Regional Ethical Review Board, Lund, Sweden, with registration number 2016/464. This ethical application permitted us to study mortality retrospectively using data from the SIR database and additionally to study laboratory findings and vital parameters obtained during admission to the general ICU at Skåne University Hospital in Lund, Sweden. All patients in SIR are entitled to have their data removed from the register or to opt not to be registered. In studies I and II, lactate and hsTnT were measured on clinical indications only.

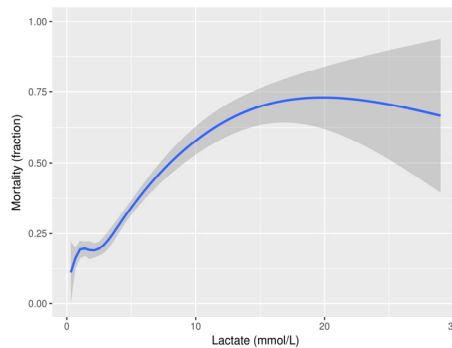
For studies IV and V, the TTM-trial protocol was approved by the ethics committees in each participating country, and informed consent was either waived or obtained from all participants or their relatives according to the national legislation, in line with the Helsinki Declaration.

# Results

## Value of additional biomarkers (studies I and II)

### Study I – The prognostic value of lactate

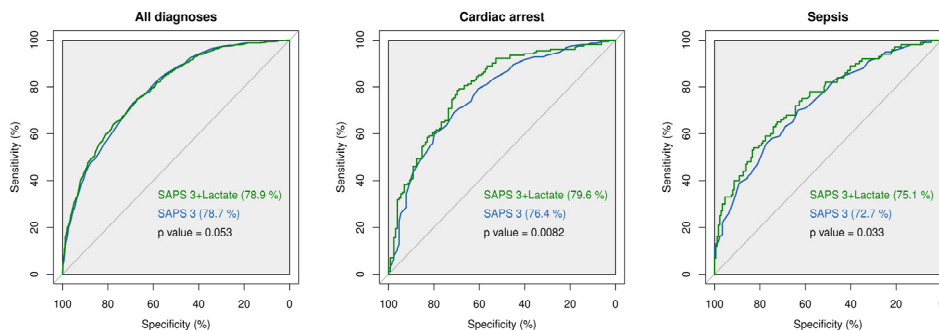
We identified 3039 patients who had their lactate concentration measured within one hour of ICU admission. Lactate was, as expected, positively correlated with 30-day mortality, as shown in figure 8.



**Figure 8. The relation between lactate levels and 30-day mortality.** 30-day mortality as a function of lactate concentration (with 95% confidence interval band) on ICU admission. Based on all 3039 patients with lactate concentration measured on ICU admission.

Using multivariate logistic regression, we found lactate to be a predictor of 30-day mortality independent of the SAPS 3 model (odds ratio [OR] 1.08, 95% CI: 1.05–1.11,  $p < 0.001$ ). Among the top five primary ICU diagnoses, we found lactate to be an independent predictor for the specific diagnoses ‘cardiac arrest’ (OR 1.17, 95% CI: 1.08–1.28,  $p < 0.001$ ) and ‘sepsis’ (OR 1.14, 95% CI: 1.05–1.25,  $p < 0.01$ ), whereas no significant results were found for ‘malignancy’ (OR 1.01, 95% CI: 0.63–1.50), ‘trauma’ (OR 1.08, 95% CI: 0.89–1.45) or ‘respiratory failure’ (OR 1.13, 95% CI: 1.91–1.42).



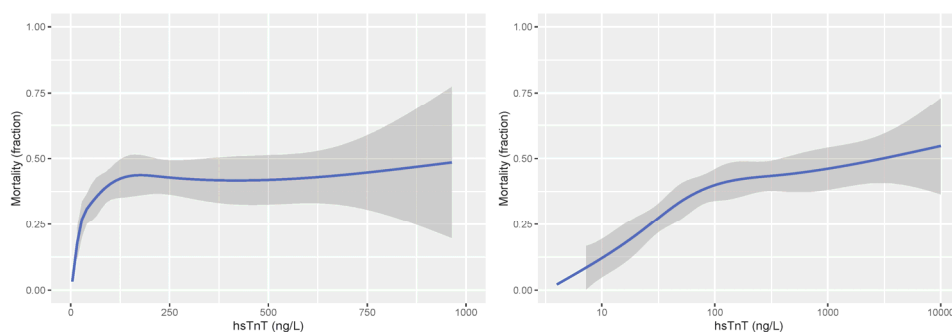


**Figure 9. Prognostic performance when adding lactate to the SAPS 3 model.** The area under the receiver operating characteristic curve (AUROC) for SAPS 3 and SAPS 3 with lactate for 'all diagnoses', 'cardiac arrest' and 'sepsis'.

As seen in figure 9, lactate did not add overall prognostic power to the SAPS 3 model as measured by the AUROC (78.9% vs 78.7%,  $p = 0.053$ ). When looking at specific diagnoses, lactate improved the prognostication for patients after cardiac arrest (AUROC 79.6% vs 76.4%,  $p < 0.01$ ) and for patients with sepsis (AUROC 75.1% vs 72.7%,  $p < 0.05$ ).

## Study II – The prognostic value of high-sensitivity troponin T

Of 4185 first-time admissions, 856 patients (20.5%) had their hsTnT measured within 90 min of ICU admission. Figure 10 shows that hsTnT was strongly correlated with 30-day mortality for hsTnT values up to 125 ng/L. For hsTnT values above 125 ng/L, the 30-day mortality remained stable at around 45%–50% up to 1000 ng/L. A more linear relationship was found between the logarithm of hsTnT up to 10,000 ng/L and 30-day mortality.



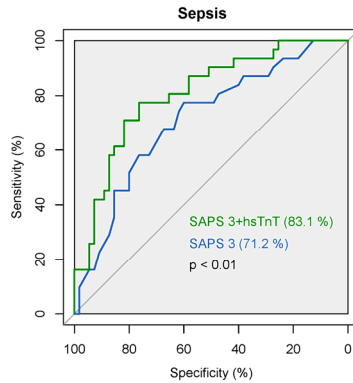
**Figure 10. The association between hsTnT values and 30-day mortality.** hsTnT values of 0–1000 ng/L are shown on the left panel to illustrate the rapidly increasing mortality rate with an increasing hsTnT value from <5 ng/L to 125 ng/L. hsTnT values <5 ng/L were replaced with 1 ng/L. On the right panel, hsTnT values up to 10,000 ng/L are shown on a logarithmic scale. The right panel is based on all 856 patients with hsTnT measured on admission. high-sensitivity troponin T: hsTnT.

Using a multivariate logistic regression on 30-day mortality, we found hsTnT to be a predictor of 30-day mortality independent of the SAPS 3 model for all diagnoses (OR 1.27, 95% CI: 1.15–1.41,  $p < 0.001$ ), for non–cardiac arrest (OR 1.37, 95% CI: 1.20–1.58,  $p < 0.001$ ) and for sepsis (OR 2.64, 95% CI: 1.63–4.75,  $p < 0.001$ ). hsTnT used as a univariate logistic regression showed good prognostic value for predicting 30-day mortality in patients with sepsis (AUROC 79.3%) and intermediate prognostic value for the overall ICU population (AUROC 65.3%) and for non–cardiac arrest patients (AUROC 68.3%). The results of the regression analyses and AUROC calculations are shown in Table 6.

**Table 6. Odds ratio and AUROC for hsTnT alone and hsTnT combined with SAPS 3 for predicting 30-day mortality.** hsTnT: high-sensitivity troponin T; AUROC: area under the receiver operating characteristic curve; SAPS 3: the 3rd version of Simplified Acute Physiology Score; CI: confidence interval. All hsTnT calculations were performed using the natural logarithm.

hsTnT alone	Odds ratio (95% CI)	p-value	AUROC hsTnT alone, %		
All	1.36 (1.25–1.49)	<0.001	65.3		
Cardiac arrest	0.97 (0.83–1.12)	0.64	52.1		
Non–cardiac arrest	1.51 (1.34–1.72)	<0.001	68.3		
– Sepsis	2.72 (1.70–4.82)	<0.001	79.3		
– Heart failure	0.94 (0.55–1.53)	0.8	53.2		
– Respiratory failure	1.13 (0.54–2.43)	0.74	51.9		
SAPS 3 & hsTnT	Odds ratio (95% CI)	p-value	AUROC SAPS 3 alone, %	AUROC SAPS 3 + hsTnT, %	p-value
All	1.27 (1.15–1.41)	<0.001	78.3	79.3	0.15
Cardiac arrest	1.07 (0.90–1.27)	0.46	78.9	78.8	0.59
Non–cardiac arrest	1.37 (1.20–1.58)	<0.001	76.1	77.6	0.16
– Sepsis	2.64 (1.63–4.75)	<0.001	71.2	83.1	<0.01
– Heart failure	0.81 (0.42–1.41)	0.48	75.3	76.8	0.57
– Respiratory failure	1.03 (0.44–2.36)	0.95	64.1	64.3	0.34

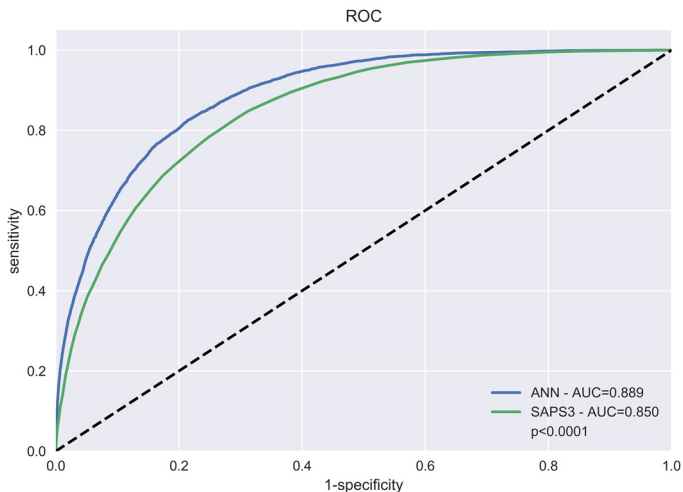
Adding hsTnT to SAPS 3 for patients with sepsis increased the AUROC by >10 percentage points (83.1% vs 71.2%,  $p < 0.01$ ), but it did not improve discrimination in the other categories. The prognostic value of hsTnT, when added to SAPS 3, for patients admitted with sepsis is shown in figure 11.



**Figure 11. Prognostic performance when adding hsTnT to the SAPS 3 model.** Comparing the areas under the receiver operating characteristic curve (AUROCs) for SAPS 3 alone and SAPS 3 with hsTnT for patients admitted to the ICU with sepsis.

## Using ANNs to predict short-term mortality (study III)

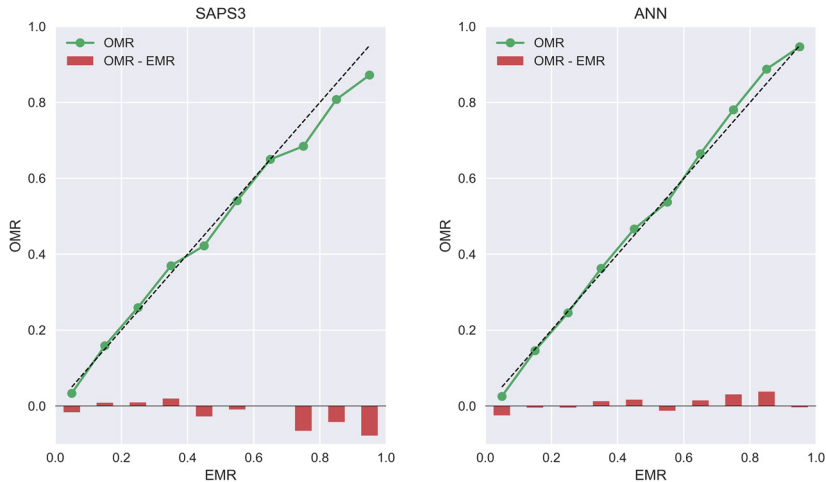
In total, 217,289 first-time admissions were identified. Of these, 181,075 patients were randomised for model development (training set), and 36,214 patients were allocated for internal validation (test set). All performance measures were based on the test set. The ANN model outperformed the SAPS 3 model (Swedish modification) according to both the AUROC (0.889 vs 0.850,  $p < 10^{-15}$ ) and the Brier score (0.096 vs 0.110,  $p < 10^{-5}$ ) in predicting 30-day mortality (see figure 12).



**Figure 12. The prognostic performance of the ANN model and the SAPS 3 model.** The area under the receiver operating characteristic curve (AUROC) for the artificial neural networks (ANN) model and the Simplified Acute Physiology Score 3 (SAPS 3; Swedish calibration 2016) model for predicting 30-day mortality. All calculations were based on the test set ( $n = 36,214$ ).

As seen in figure 13, the calibration error (the difference between the OMR and EMR) in the high-EMR range was reduced in the ANN model.

The performances of the ANN model and the SAPS 3 model for different primary diagnoses can be compared in table 7. The ANN model outperformed the SAPS 3 model for all top primary diagnoses.



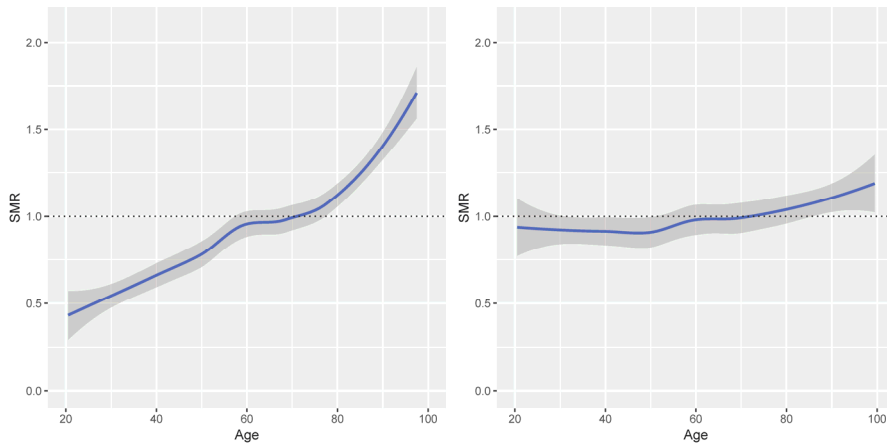
**Figure 13. Calibration curves.** Observed mortality rate (OMR) vs estimated mortality rate (EMR). On the left, EMR is calculated based on the Swedish calibration of the Simplified Acute Physiology Score (SAPS 3) from 2016. On the right, EMR is calculated based on the ANN model. Note the difference between OMR and EMR for the SAPS 3 model in the high-EMR range. All calculations were based on the test set ( $n = 36,214$ ).

**Table 7. Prognostic capability for different primary ICU diagnoses.**

The performance of the Simplified Acute Physiology Score (SAPS 3) model and the artificial neural network (ANN) model for different primary ICU diagnoses. All calculations were based on the test set ( $n = 36,214$ ). The area under the receiver operating characteristic curve (AUROC) is presented, with a 95% confidence interval in brackets. SIRS: Systemic Inflammatory Response Syndrome.

	Number of patients	AUROC of SAPS 3	AUROC of ANN	p-value
Test set	36,214	0.850 (0.846–0.855)	0.889 (0.885–0.893)	$<10^{-15}$
Cardiac arrest	1,651	0.858 (0.835–0.881)	0.893 (0.875–0.912)	$<10^{-7}$
Septic shock	1,481	0.846 (0.821–0.870)	0.889 (0.869–0.909)	$<10^{-8}$
Respiratory failure	1,467	0.830 (0.804–0.856)	0.878 (0.855–0.900)	$<10^{-8}$
Gastrointestinal haemorrhage	1,324	0.878 (0.858–0.900)	0.910 (0.892–0.927)	$<10^{-5}$
SIRS	1,320	0.836 (0.811–0.862)	0.884 (0.863–0.906)	$<10^{-8}$
Trauma	1,301	0.844 (0.820–0.869)	0.882 (0.860–0.903)	$<10^{-5}$
Bacterial pneumonia	1,173	0.856 (0.830–0.882)	0.895 (0.874–0.916)	$<10^{-7}$
Seizures	797	0.847 (0.814–0.880)	0.892 (0.865–0.918)	$<10^{-4}$
Head injury	760	0.833 (0.796–0.869)	0.888 (0.860–0.916)	$<10^{-5}$

After ranking all SAPS 3 variables and then adding one variable at a time to the model, we found that using ANNs to interpret eight variables produced a model that achieved performance similar to that of the SAPS 3 model. The eight variables were (in order of importance for improving the AUROC) age, level of consciousness, neurological cause, cardiovascular cause, cancer, temperature, pH and leukocytes. The eight-variable model had an AUROC of 0.851 (95% CI: 0.845–0.857) and a Brier score of 0.106 (95% CI: 0.106–0.107).



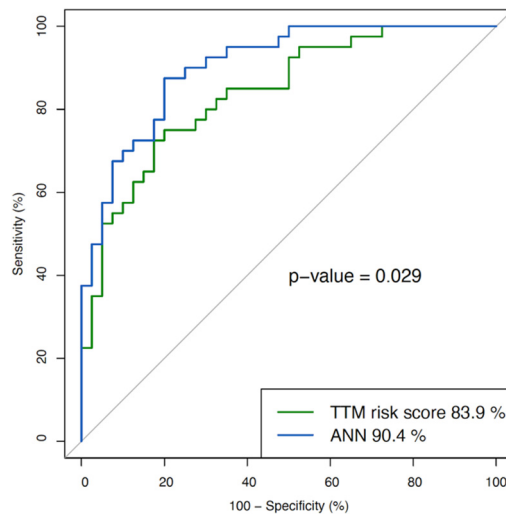
**Figure 14. Age and the standardised mortality ratio (SMR).** The SMR is the observed mortality rate (OMR) divided by estimated mortality rate (EMR): an OMR above 1 represents higher mortality than expected and vice versa. The SMR is displayed as a function of age (which is the single most prognostic factor in SAPS 3) for the Simplified Acute Physiology Score (SAPS 3) model (left panel) and the artificial neural network (ANN) model (right panel) for the test set ( $n = 36,214$ ). SMR is shown with a 95% confidence interval.

As can be seen in figure 14, the ANN model was superior to the SAPS 3 model in correcting for age, which is the most important variable in SAPS 3. The SAPS 3 model underestimated 30-day mortality in the elderly ICU population and overestimated 30-day mortality in the younger ICU population.

# Predicting neurological outcome after out-of-hospital cardiac arrest (studies IV and V)

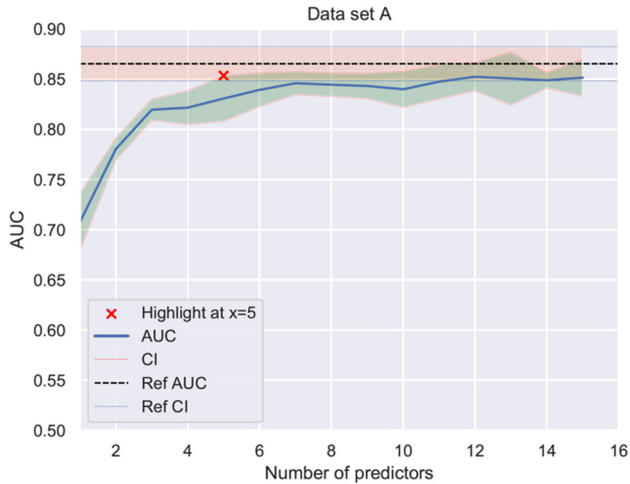
## Study IV – Cardiac arrest prognostication on admission

In study IV, a total of 932 patients and 54 variables were included for final analysis. We randomly selected 839 patients (90%) for model development (training set) and 93 patients (10%) to evaluate the model's prognostic performance (test set). The cross-validated AUROC (for the training set) was 85.2%, and the AUROC when evaluating performance using the test set was 89.1%.



**Figure 15. Performance comparison of the TTM risk score and the ANN model.** Comparison between the performance of the TTM risk score and our ANN model based on the 80 patients in the test set. TTM: targeted temperature management; ANN: artificial neural network.

We compared the performance of our ANN model with that of a similar prediction model, the ‘TTM risk score’ as described in the Background chapter (74). The ANN model performed significantly better (AUROC 90.4% vs 83.9%,  $p = 0.029$ ), as shown in figure 15. Note that 13 patients were removed from the test set before comparing the two models, as the TTM risk score model could not handle missing values. This change in the test set explains the difference between the ANN model AUROC determined here of 90.4% and the 89.1% reported above.



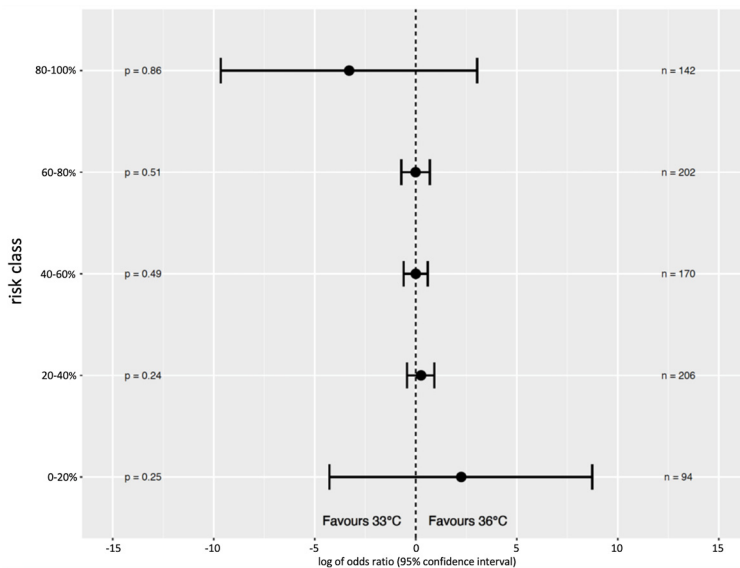
**Figure 16. Model performance as variables are added.** In the search for a simplified model, we ranked all input variables by subtracting one variable at a time from the developed model and calculating the AUROC. We then started with the most important variables from this ranking and added one variable at a time back to the model, calculating the AUROC at each step based on the training set. The AUROC (AUROC = AUC in figure 16) is based on the cross-validation (training), represented by the blue line with its corresponding CI. The best performing model, with 54 variables, is represented by the dotted line with its corresponding CI. The CIs overlapped after five variables had been added; this is marked with a red X to indicate the point after which no significant difference was found between the two models. AUROC: area under the receiver operating characteristic curve; CI: confidence interval.

**Table 8. Model performance as variables are added.** All variables were ranked and then added one at a time, starting with age, to build a model from the ground up. The model performance is shown as the AUROC with the corresponding CI based on the cross-validation (training; see figure 16), and the corresponding test set performance is also shown. For comparison, the performance of the final model with all 54 variables is shown as well. AUROC: area under the receiver operating characteristic curve; CI: confidence interval; ROSC: return of spontaneous circulation; GCS: Glasgow Coma Scale; AMI: acute myocardial infarction; CV: cross-validation.

No. of variables	Variables	AUROC <sub>CV</sub>	AUROC <sub>test</sub>
1	Age	0.708 (±0.0286)	0.657
2	+ Time to ROSC	0.780 (±0.0113)	0.799
3	+ First monitored rhythm	0.820 (±0.0106)	0.852
4	+ Previous cardiac arrest	0.822 (±0.0169)	0.861
5	+ GCS motor score	0.832 (±0.0229)	0.863
6	+ Dose of adrenaline	0.839 (±0.0170)	0.826
7	+ Creatinine	0.846 (±0.0117)	0.837
8	+ Cardiac arrest location	0.854 (±0.0119)	0.857
9	+ Previous AMI	0.843 (±0.0129)	0.835
10	+ Diabetes	0.840 (±0.0182)	0.844
11	+ Length	0.848 (±0.0173)	0.869
12	+ Time to Advanced CPR	0.853 (±0.0142)	0.870
13	+ pH	0.851 (±0.0266)	0.880
14	+ Platelets	0.849 (±0.0079)	0.875
15	+ Bystander witnessed arrest	0.852 (±0.0188)	0.886
54	All variables	0.852 (±0.0172)	0.891

We ranked all 54 variables based on the size of the effect on the AUROC when they were removed from the model. We then added one variable at a time, starting with the most important variable according to the ranking, and calculated the AUROC based on the training set at each step. The AUROC calculations for the 15 most important variables are shown in figure 16 and described in further detail in table 8.

After hospital admission, patients in the TTM-trial were randomised to a target temperature of 33°C or 36°C for 28 h. By dividing the cohort into five risk groups based on the ANN model predictions, we could investigate whether one of the two treatments would benefit certain risk groups more than others. We found that no specific risk group benefitted from a specific target temperature, as shown in figure 17.

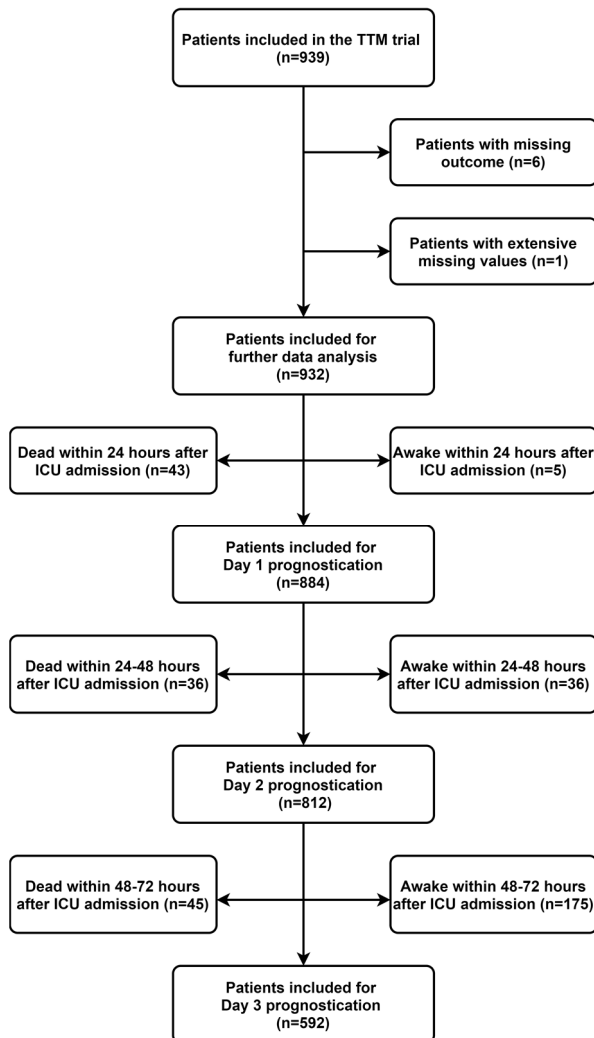


**Figure 17. Treatment effect based on the ANN-model-stratified risk groups.** Patients were divided into five groups based on their probability of a poor outcome at hospital admission using the ANN model. An odds ratio over 1 indicates a better functional outcome when treated with 36°C compared to with 33°C and vice versa.



## Study V – Cardiac arrest prognostication during the first three days after ICU admission

As in study IV, 932 patients were included for further analysis after six patients had been removed due to missing outcomes and one patient due to missing values. As detailed in figure 18, patients who had either awakened or died during the time windows 0–24 h (day 1), 24–48 h (day 2) and 48–72 h (day 3) after ICU admission were removed accordingly.

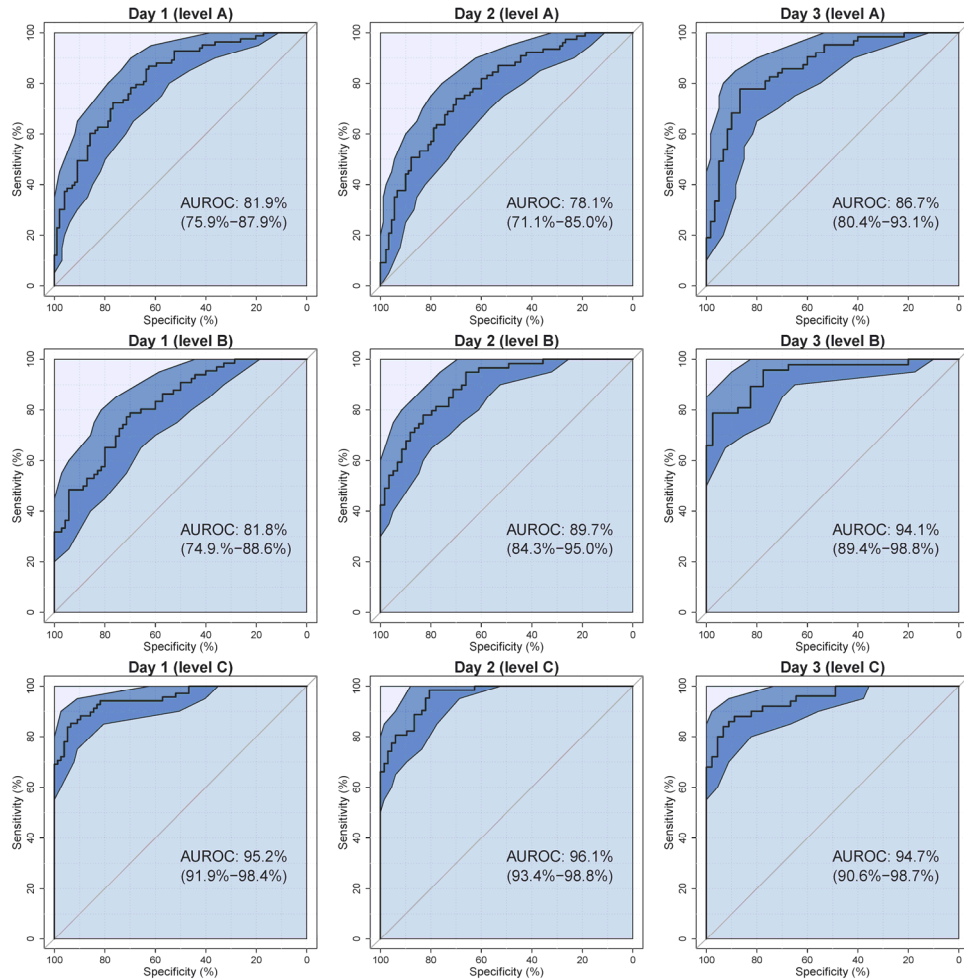


**Figure 18. Flowchart.** Flowchart for the study populations at days 1 (24 h), 2 (48 h) and 3 (72 h) after ICU admission. After each day, patients who had died or awakened were removed from further analysis to focus on the prognostication of comatose patients. Furthermore, on each day, three datasets were created that reflected the levels of biomarkers included in the model (not shown here). Population characteristics are based on 'Patients included for further data analysis (n = 932)' (see table 1A–D under supplements in study V). TTM: targeted temperature management; ICU: intensive care unit.

**Table 9. Model overview and prognostic performance.** In level A, we used all clinical variables available from the TTM-trial. In level B, we added clinically accessible biomarkers. For level C, we added research-grade biomarkers as well. The prognostic performance is displayed as the AUROC and using a confusion matrix. Note that the confusion matrix threshold was based on the threshold for 100% specificity in the training set. AUROC: area under the receiver operating characteristic curve; TN: true negative; TP: true positive; FP: false positive; FN: false negative.

Timeline	Type of data	Number of patients		Number of variables		Model performance		Confusion matrix (test set)					
		Total (n)	Train set (n)	Test set (n)	Before variable selection (n)	After variable selection (n)	Training set (cross validation) AUROC (%) (CI 95 %)	Test set AUROC (%) (CI 95 %)	Probability threshold	TN (n)	FP (n)	FN (n)	TP (n)
Day 1 (24h)	Level A	884	702	182	120	22	85.7 (83.2-88.6)	81.9 (75.9-87.9)	0.981	99	0	82	1
	Level B	638	502	136	125	21	89.9 (86.8-92.2)	81.8 (74.9-88.6)	0.983	70	0	57	9
	Level C	690	545	145	131	22	96.6 (94.7-97.5)	95.2 (91.9-98.4)	0.887	76	1	21	47
Day 2 (48h)	Level A	812	645	167	174	21	85.8 (82.9-88.5)	78.1 (71.1-85.0)	0.935	88	2	70	7
	Level B	578	460	118	184	17	93.7 (91.7-95.9)	89.7 (84.3-95.0)	0.963	59	0	44	15
	Level C	624	495	129	196	21	96.6 (95.2-97.9)	96.1 (93.4-98.8)	0.986	67	0	31	31
Day 3 (72h)	Level A	592	469	107	228	17	84.9 (80.6-87.9)	86.7 (80.4-93.1)	0.920	60	0	22	25
	Level B	415	328	87	243	12	92.7 (89.8-95.1)	94.1 (89.4-98.8)	0.885	40	0	22	25
	Level C	442	347	95	261	23	96.6 (94.9-98.1)	94.7 (90.6-98.7)	0.820	45	0	16	34

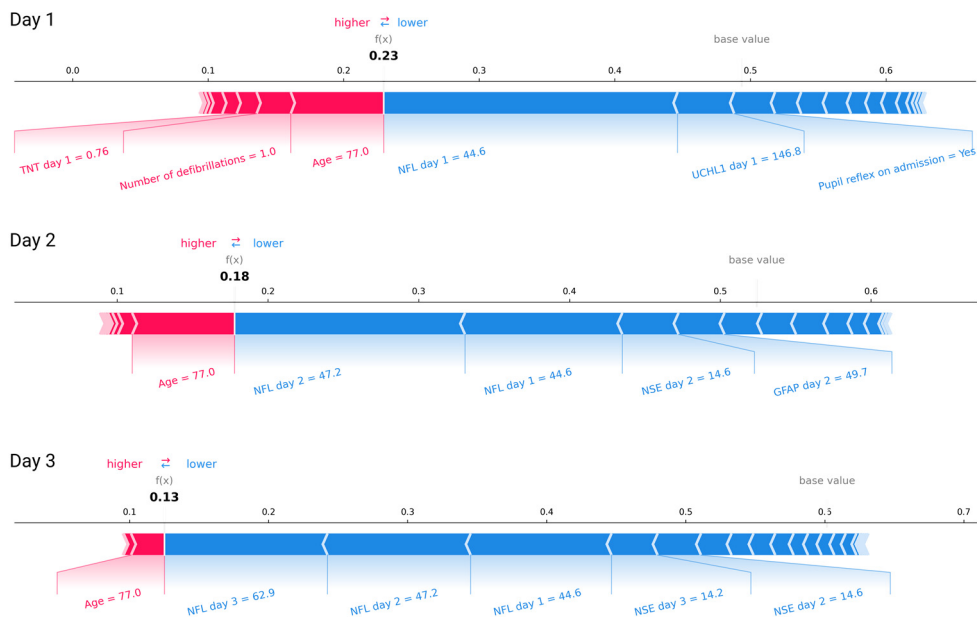
As described in the Methods and materials chapter, for each day, three datasets were created that included three different levels of biomarkers. The datasets varied based on the number of patients with missing values. The variable counts before and after variable selection and the number of patients in each dataset are given in table 9.



**Figure 19. Prognostic performance for all nine models.** Model performance predicting poor neurological outcome after six months based on the corresponding test set. The columns represent the timeline after ICU admission. The rows represent the different levels of biomarkers added to the available clinical variables from the TTM-trial: none (level A), clinically accessible biomarkers (level B) and research-grade biomarkers (level C). The CI for the AUROC is calculated for each model. Furthermore, the CI for the specificity at different levels of sensitivity is displayed as a blue CI band. AUROC: area under the receiver operating characteristic curve; CI: confidence interval.

ROC curves indicating prognostic performance are illustrated in figure 19. For the level A model, without added biomarkers, prognostic performance remained under 90% during the three days after ICU admission. For the level B model, which included accessible biomarkers, the prognostic performance improved significantly from day 1 to day 3 (from 81.8% to 94.1%,  $p < 0.01$ ). For the level C model, with research-grade biomarkers included, the performance was excellent from day 1 through day 3.

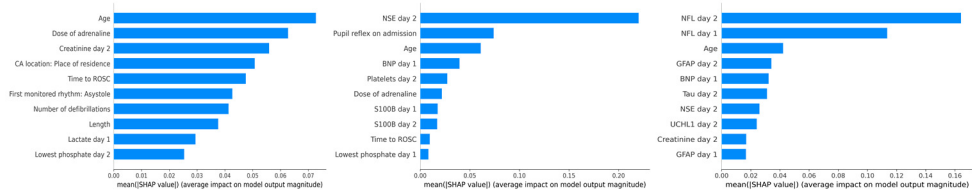
A notable finding was that level C at all time points, and level B at 72 h, had a sensitivity above 60% while retaining a 100% specificity. When we investigated this further, by using a threshold of 100% specificity (in the training set) to predict the outcome in patients within the test set, two models generated false-positive predictions (predicted poor outcome, reported good outcome). Most of the models had a high rate of false-negative predictions (predicted good outcome, reported poor outcome), but that rate remained under 25% when using research-grade biomarkers (level C). Model performance details are shown in table 9.



**Figure 20. The Shapley additive explanations (SHAP) algorithm used to explain how patient-specific predictions were generated.** The patient in this example was predicted to have a risk of a poor outcome of 23% on day 1, 18% on day 2 and 13% on day 3 (using the level C model). On all three days, the patient's age was the factor that contributed most to increasing the risk, and modest levels of NFL and NSE contributed most to decreasing the level of risk overall. TNT: troponin-T (ng/L); NFL: neurofilament light (ng/L); UCHL1: ubiquitin carboxy-terminal hydrolase L1 (ng/L); NSE: neuron-specific enolase (ng/mL); GFAP: glial fibrillary acidic protein (ng/L).

The SHAP algorithm was applied to all models to explain how each prediction had been generated. A patient example is illustrated in figure 20, where predictions of the ANN model that uses research-grade biomarkers (level C) are explained using the SHAP algorithm on days 1, 2 and 3. The patient’s age adds to an increase in the risk of poor outcome, while the low levels of the biomarkers decrease the risk.

Moreover, we used the SHAP algorithm to rank all variables for each model, as illustrated in figure 21, which displays all three levels for day 2. As seen in this figure, the list of the top 10 most important variables changes when clinically accessible and research-grade biomarkers are added. Age and the dose of adrenaline given are the two most important variables in level A. The importance of these variables is reduced in levels B and C, and in level C the dose of adrenalin is not included in the top 10 variables, while age remains the third most important variable.



**Figure 21. SHAP variable importance on day 2.** The global importance of each variable in each model illustrated with the SHAP variable importance. The most important variable has the highest mean of absolute SHAP values. The left panel shows level A, the middle panel shows level B (adding clinically accessible biomarkers), and the right panel shows level C (adding research-grade biomarkers). Similar figures for day 1 and 3 can be found in study V under supplements. CA: cardiac arrest; ROSC: return of spontaneous circulation; NSE: neuron-specific enolase; BNP: brain natriuretic protein; S100B: S100 calcium-binding protein B; NFL: neurofilament light; GFAP: glial fibrillary acidic protein; UCHL1: ubiquitin carboxy-terminal hydrolase L1.

# Discussion

The most commonly used ICU scoring systems have continuously developed, along with intensive care medicine, over the last four decades. This thesis is another step in the search for better prognostication for our most critically ill patients. We first investigated the effects of adding single biomarkers to our current scoring system and then focused on using ANNs to develop prediction models for both critically ill patients in general and in comatose patients post-OHCA. This chapter merges the overall points from the discussions in the five studies.

In the first two studies, we used rather simple techniques to investigate whether inclusion of a biomarker added value to the current SAPS 3 model. We removed patients with missing values and used logistic regression on 30-day mortality and found significant associations (odds ratios) for both lactate and hsTnT. Even though the odds ratios were significant, neither lactate nor hsTnT added prognostic value to SAPS 3 as measured by the AUROC. We were, however, able to show added prognostic value in sepsis and in cardiac arrest (lactate only).

In the third study, we developed an ANN using the SAPS 3 variables. The results of the ANN were distinctly better than the SAPS 3 model, especially the way the ANN model corrected for age (the single most important prognostic variable). We were also able to design an ANN using only eight SAPS 3 variables that provided performance similar to that achieved by SAPS 3. ANNs also showed promising results in study IV. Even though the number of patients represented in the data used for model development was much lower, the ANN model outperformed the TTM risk score – a logistic regression-based model.

Studies IV and V both predicted long-term neurological outcome post-OHCA based on either admission data (study IV) or data obtained during the first three days after ICU admission (study V). The black-box nature of ANNs was addressed in study V with the SHAP algorithm. It is important to note that the SHAP algorithm, powerful as it is, does not improve the model performance; rather, it explains the reason behind each prediction. Studies III and V in particular illustrate how machine learning could aid intensive care medicine in the future.

## Value of additional biomarkers (studies I and II)

We investigated the predictive capabilities of lactate and hsTnT in studies I and II. Although both biomarkers were found to be associated with 30-day mortality independently of the SAPS 3 model for all diagnoses, the prognostic value of the biomarkers was not sufficient to impact the model's overall AUROC. When investigating the prognostic performance for specific diagnoses, however, we found both biomarkers added substantial value to SAPS 3 for patients admitted with sepsis. Lactate also added value to SAPS 3 for patients admitted after cardiac arrest. The greatest effect was found upon adding hsTnT to SAPS 3 for patients with sepsis, improving the SAPS 3 AUROC by more than 10% (from 71.2% to 83.1%,  $p < 0.01$ ). Even as a single biomarker, hsTnT had an AUROC of 79.3% for patients with sepsis.

Other studies investigating TnT have also found similar effects for critically ill patients with sepsis (40, 123). However, Røsjø et al. (124) did not find that the hsTnT level at admission added value to their scoring system (SAPS 2, the predecessor to SAPS 3), and hsTnT alone had a lower AUROC than reported in our study. Our study design had limitations which could explain some of the differences between our findings and those of Røsjø et al.'s study. In our study, patients had hsTnT measured only on clinical indication, making them a subgroup of sepsis patients with a clinical suspicion of cardiac injury. The EMR was significantly higher in the group that had hsTnT measured (hsTnT group) compared to the group that did not (non-hsTnT group).

Moreover, we did not have ECG records which could have clarified the cause of the elevated hsTnT. Despite the limitations of study II, the prognostic value of hsTnT as a biomarker for patients with sepsis is noteworthy and should be investigated further in a prospective study. Another possible weakness (of both studies I and II) is the possible variance among physicians when deciding the primary diagnosis for the ICU admission, as this could add noise to the model. Lengquist et al. found large discrepancies in diagnosing sepsis as a primary diagnosis compared to using the Sepsis 3 criteria in a retrospective comparison (125).

It is important to find strong predictors of ICU mortality to simplify ICU prognostication and improve the prognostication for specific diagnoses. However, for future studies, it would be better if the biomarkers, such as lactate or hsTnT, were included in the model development, so that they can compete on equal terms with the other variables included in the model. This requires prospective studies and larger datasets to avoid the limitations mentioned here. Furthermore, an external test set should be used to validate any findings. Neither study I nor study II used an independent test set.

## Using ANNs to predict short-term mortality (study III)

In study III, we investigated whether the SAPS 3 prognostication could be improved by using ANNs trained on the SAPS 3 variables. We found ANNs to be superior to the SAPS 3 model (Swedish modification) in predicting 30-day mortality measured by the AUROC and the Brier score. A 4% increase in the AUROC from 85% to 89% may not sound substantial, but as the AUROC approaches a perfect score of 100%, this degree of improvement becomes increasingly difficult. As an example, when we analysed how well-calibrated age (the single most important predictive variable) was in both models, the difference between the ANN model and the SAPS 3 model become apparent (see figure 14).

The SAPS 3 model is a logistic-regression-based model which has been calibrated over time using three parameters, as described in the Background chapter. This means that SAPS 3, powerful as it is, has a rigid structure which theoretically could have drawbacks when variable interactions are complex. In comparison, the ANN model developed in study III, a non-linear model, incorporated more than 10,000 weights (adjustable parameters). A direct comparison of the number of parameters used by the two models, however, is of limited use, as the SAPS 3 model also uses a predefined scoresheet when calculating the SAPS 3 score, which is then converted into the EMR. All things considered, when the two methods were compared, we found the ANN approach to be superior in predicting 30-day mortality. Our results, however, differ from those of a recent systematic review, which did not find machine learning models superior to logistic regression models (126).

To our knowledge, only one other study has applied similar AI methods to predict mortality based on SAPS 3 variables. Thorsen-Meyer et al. used a subset of the SAPS 3 variables with a recurrent neural network to predict 90-day mortality in a Danish ICU cohort (62). Their predictive performance measured by AUROC on admission was around 75%. As a comparison, Engerström et al. found the SAPS 3 model could predict 90-day mortality in a Swedish population with an AUROC of 84% (26). However, as their study's variables, study population, and outcome differed from those used in study III, a direct comparison of the studies is meaningless. Nevertheless, the difference between Thorsen-Meyer et al.'s findings and ours seems larger than expected.

When developing prediction models, there is often a threshold between complex yet high-performing prediction models and simpler models with more moderate performance. For each model, this trade-off between simplicity and performance has to be considered. In study III, we developed a simple eight-variable model with performance similar to that of SAPS 3, which underlines the strength of ANNs.

Study III had two main limitations which could be improved on in future studies. If the test set had been comprised of data from the latest year (e.g. 2017 in study III) instead of data chosen completely at random, the performance measure would



reflect a more recent patient population. Second, as with all ANN models, there are black-box concerns, meaning that it is very difficult to understand how the model works. Applying the SHAP algorithm after model development can reduce some of these black-box concerns. This is discussed later in this chapter.

When we investigated specific diagnoses, the ANN model was superior for the most common primary diagnoses, from cardiac arrest to septic shock. The ANN model achieved a near-excellent AUROC of 89.3% when predicting for patients admitted after cardiac arrest (non-specific).

## Predicting neurological outcome after out-of-hospital cardiac arrest (studies IV and V)

While study IV focused on prognostication at admission to the hospital, study V focused on daily prognostication until day 3 (72 h), the earliest time point to start prognostication according to the current guidelines (78). The designs of these two studies could give insight into how these models' prognostic performance evolves during the first 72 h after ICU admission. Both study IV and study V are post hoc analyses of the TTM-trial, which enrolled only patients admitted to the hospital after OHCA of presumed cardiac cause. Moreover, the study had several exclusion criteria (such as limitations in therapy, including do-not-resuscitate orders, or known illness making survival to 180 days unlikely) which must be considered when interpreting the results.

In study IV, the ANN model using 54 variables outperformed the TTM risk score, a logistic-regression-based model developed on the same dataset. Even when the ANN model used only three to five variables, its prognostic performance was good when evaluated on the test set. This leads to the question of how many variables we should include in our models. In both study III and study IV, we primarily used all accessible variables and secondarily created a simplified model. These two models had different aims; the simplified model is easy to use bedside but is not as accurate as the model with more variables, which is better suited for use as an integrated part of electronic medical records. In study V, we used a variable selector to reduce the number of variables for each model, in some cases by 90%. Which method to choose depends on the situation, the prognostic strength of the biomarkers and the aim of the prediction. Figure 16 (study IV) illustrates this.

Study V showed how important research-grade biomarkers such as NFL are, reporting an AUROC of around 95% only one day (24 h) after ICU admission. This prognostic performance remained at the same level after day 2 (48 h) and day 3 (72 h). These findings correspond well with previous findings; the only difference between those earlier studies and ours is that our study focused solely on comatose

patients (88, 98). When only clinically accessible biomarkers such as NSE were added, the prognostic performance improved significantly from day 1 to day 3. When using clinically accessible biomarkers on day 3, the performance was similar to the performance using research-grade biomarkers (level C) on all three days. Analysing the ROC curves reveals that these four models had performances comparable to that of the ERC–ESICM guidelines as reported by Moseby-Knappe et al. (81). This is interesting, as our models did not have access to crucial prognostic information such as EEG, SSEP or neurological imaging. When we investigated this even further by using a threshold based on 100% specificity on the training data (cross-validations) to predict the outcome in the test sets, the model with research-grade biomarkers (level C) on day 1 (24h) had one false-positive prediction.

The ERC–ESICM guidelines are used only for a subset of patients on day 3 (after 72 h), as the patients must have a GCS-M  $\leq 2$  before the prognostication can be initiated. Otherwise, the prognostication will be postponed. The setup in study V, by contrast, created predictions for all comatose patients.

It is essential to acknowledge the uncertainty present when using datasets that are the size of the TTM-trial dataset. The performance of models using such datasets can be too dependent on the train/test set split and vulnerable to outliers in general. For example, the model performance was presumably better at the time of hospital admission (study IV) than after 24 h without the use of biomarkers (study V). This difference is noteworthy and must be kept in perspective when discussing this approach to cardiac arrest prognostication.

Both studies IV and V used a dichotomised CPC score as the outcome of the prediction models. Simplified as it is, this score gives information about the neurological outcome. Optimally the model would instead predict the specific CPC score. However, given the size of the TTM-trial dataset, an attempt to predict a specific score would probably lead to prediction uncertainties. Our hope is that data from the 1900 patients enrolled in the TTM2-trial will soon make such multiclass predictions meaningful and allow valuable predictions to be made at different time points: mRS and CPC score at ICU discharge, hospital discharge, after six months and after 24 months.

## Improving ICU prognostication

The ICU is a very data-rich environment. For the most critically ill patients, hundreds of different variables are gathered daily, from basic vital parameters, laboratory findings and various radiological imaging to continuous information streams such as ECGs, EEGs and arterial pressure curves in high resolution. Moreover, information from the technical equipment used in the ICU, such as ventilators, infusion pumps (medicine and fluid infusion) and dialysis machines,

still largely remains unused in prognostication. Today, the ICU physician interprets this highly complex matrix of information to plan the best possible treatment for the individual patient. The goal is to ensure the patient will receive state-of-the-art care, survive and get safely through their ICU stay without treatment complications. AI algorithms could support ICU physicians in reaching this goal by continuously predicting the risk of mortality and morbidity as the critically ill patient is being treated.

In recent years, numerous publications have focused on predictive modelling for critically ill patients, some focusing on the deterioration in the patient's condition outside the ICU while others focus on patients admitted to the ICU. Lauritsen et al. achieved high predictive performance using a machine learning algorithm to predict acute critical illness based on electronic health records (127). Focusing on patients after admission to the ICU, several studies have shown how predictive performance improved over the course of an ICU stay (61, 62, 128). These studies incorporated modern machine learning algorithms and data from electronic health records into their predictions.

The next natural step for ICU prognostication will be to follow critically ill patients before any possible ICU admission and after discharge to ensure the best possible long-term outcome. As shown above, this is possible using modern computer algorithms. The challenges will be in handling missing values and noise in a sophisticated manner and in the subsequent implementation. In this thesis, missing values were either excluded, as in studies I and II, or imputed using simple median or mode substitution for continuous and binary variables, respectively. As ANNs cannot handle missing values by themselves, it is important to use imputation methods to replace missing values. During model development, more advanced imputation methods such as autoencoders (study III) and missforest (study V) did not improve model performance. If imputation techniques are needed, we recommend a grid search of imputation methods to find the best possible one.

Sixty years since the establishment of the 'first ICU prototype' at the University of Southern California and we have barely started to integrate AI methods into the care of critically ill patients. We are in a transition period in which our surroundings are utilising data in a manner never seen before, yet this revolution has not yet fully reached the healthcare system. There is a need to further develop these methods and move them into our clinical practices for the future well-being of our patients.

# Conclusions

Lactate was found to be an independent predictor of 30-day mortality in addition to SAPS 3. The addition of lactate to SAPS 3 improved the AUROC for patients admitted with sepsis or cardiac arrest as their primary diagnoses, although not for all diagnoses.

High-sensitivity TnT (hsTnT) was also found to be an independent predictor of 30-day mortality when added to SAPS 3. For patients admitted with sepsis as their primary diagnosis, hsTnT improved the prognostic performance measured by AUROC by more than 10%. Further studies are needed to validate the strength of hsTnT in sepsis prognostication.

By using ANNs to interpret the variables used in the SAPS 3 model the prognostic performance improved noticeably. The ANN model outperformed the SAPS 3 model (Swedish modification), measured by the AUROC, the Brier score and calibration plots, in predicting 30-day mortality when evaluating the model based on more than 36,000 patients (internal validation). The ANN model was superior to SAPS 3 in correcting for age. Furthermore, an ANN model developed using only eight variables showed similar performance as the full SAPS 3 model.

For comatose patients admitted to the ICU post-OHCA, our ANN model was superior to a logistic-regression-based model in predicting the neurological outcome based on information available at hospital admission. No specific risk groups benefitted from a TTM of neither 33° C nor 36 ° C.

When using clinical variables with and without clinically accessible and research-grade biomarkers during the first three days after ICU admission, ANN models showed good to excellent prognostic performance in predicting the neurological outcome in comatose patients post-OHCA. Especially, the models which included NSE after 72h and NFL on all days showed promising prognostic performance.



# Future perspectives

## Continuing to improve early ICU prognostication

SAPS 3 uses background information and information obtained within the first hour after ICU admission. This makes SAPS 3 a useful tool for benchmarking and research. We should utilise the prognostic information available from other healthcare registries on information about previous diagnoses, prescriptions and so on. Moreover, data science is constantly evolving; new methods for imputing missing values, new feature selection methods and new supervised machine learning algorithms should continuously be tested to improve early ICU prognostication.

## Developing dynamic models

On a regional level, it is possible to create and implement dynamic models, based on time-series data, that continuously predict the patient's risk of deterioration. The time resolution can range from daily for variables such as biomarkers to milliseconds for variables such as ECG and EEG. A complicating factor is how to utilise the national registries information mentioned above so that the prognostic performance for patients at ICU admission would be comparable with national standards. Otherwise, training AI models on smaller datasets could result in significantly lower performance.

## Morbidity prognostication

After development, these prediction models (early and dynamic models) can relatively easily be retrained to predict various morbidity indicators (direct or indirect measures) such as kidney failure, depression or level of assistance provided by the government. In this way, we can create models for other short-term and long-term outcomes after ICU admission in addition to the standard mortality predictions. This would be valuable information for physicians, researchers, patients and patients' family.

## Prediction on an individual level

All of the above goals require large datasets to be able to give meaningful predictions on an individual level. The current ICU scoring systems are regarded as too uncertain to be used for individual predictions. To be able to use future prediction models on an individual level, we should test methods which can detect outliers or can, in another way, explain the uncertainty in each prediction.

## Improving post–cardiac arrest prognostication

With almost 2000 post-OHCA patients, the upcoming TTM2-trial (the sequel to the TTM-trial) database could be used to overcome some of the limitations of studies IV and V. Hopefully, we will be able to improve prognostic performance by adding prognostic information such as EEGs, SSEP and neuroimaging to our models to predict specific mRS or CPC scores.

# Populärvetenskaplig sammanfattning

På intensivvårdsavdelningar (IVA) vårdas patienter med akuta och livshotande tillstånd såsom svår blodförgiftning, efter trafikolyckor eller vård efter hjärtstopp. Patienter på intensivvårdsavdelningar kräver noggrann observation och behandling dygnet runt. Behandling på IVA består i att stötta eller ersätta vissa organfunktioner tills patientens tillstånd förbättras eller en specifik behandling har haft effekt. Avhandlingen handlar om hur riskbedömningen av kritiskt sjuka patienter som läggs in på en intensivvårdsavdelning kan förbättras genom att inkludera information från blodprover och med hjälp av artificiell intelligens (AI).

Trots moderna behandlingsmetoder så har patienter på intensivvårdsavdelning en hög dödlighet - faktiskt dör en av sex inom 30 dagar från inläggning. För vissa patientgrupper, t.ex. patienter som är medvetslösa efter ett hjärtstopp, så är dödligheten hela 50%. För att kunna följa upp effekten av behandlingar så behövs ett bra verktyg för riskbedömning vid inläggning på IVA. Den modell som används i Sverige idag heter the Simplified Acute Physiology Score 3 (SAPS 3). Modellen använder sig av information om patientens medicinska bakgrund, inläggningsorsak, fysiologiska mätvärden som t.ex. blodtryck och blodprover som avspeglar olika organfunktioner. Med hjälp av de här värdena räknar den ut risken för att patienten avlider inom 30 dagar. Sedan modellen skapades 2005 så har den grundläggande strukturen varit densamma och inga nya blodprover har lagts till för att försöka förbättra riskbedömningen.

I de första två studierna undersökte vi om två olika blodprover kunde tillföra viktig information till SAPS 3-modellen. De två blodproverna var mjölksyra (laktat) och "högekänsligt troponin T" (hsTnT) som används för att påvisa blodpropp i hjärtat. Vårt resultat visar att laktat och hsTnT förbättrar riskbedömningen för patienter med sepsis (blodförgiftning) och hjärtstopp (enbart laktat).

Förbättrade algoritmer och större datorkraft har ökat möjligheterna att använda AI inom medicinsk forskning. I denna avhandling användes en AI-metod som heter artificiellt neuralt nätverk (ANN). Förenklat kan man säga att metoden på digital väg försöker efterlikna biologiska nätverk med nervceller (som i hjärnan), där enskilda "celler" tar emot och skickar information till varandra. Denna struktur kan vara en fördel när man ska hitta komplexa samband mellan många olika variabler såsom ålder, blodprover, samsjuklighet och utfall som t.ex. död eller dålig funktionsnivå.



I den tredje studien undersökte vi om just ANN kan förbättra riskbedömningen vid patientens ankomst till IVA genom att använda sig av samma variabler som i SAPS 3-modellen. Vi använde anonymiserade data från mer än 180 000 patienter till att bygga vår modell och 36 000 patienter för att utvärdera hur bra den fungerar. Med hjälp av ANN blev riskbedömningen av intensivvårdspatienter i Sverige bättre på alla sätt vi undersökte. Neurala nätverk kan bland annat hantera patientens ålder på ett bättre sätt, då äldre och yngre patienters dödlighet under- respektive överskattats av SAPS 3-modellen. Metoden var dessutom bättre i riskbedömningen av alla de vanligaste diagnoserna på intensivvårdsavdelningar.

Patienter som överlever hjärtstopp utanför sjukhuset och är i koma efteråt utgör en speciell patientgrupp på intensivvårdsavdelningen. Ofta ses en komplex sjukdomsbild p.g.a syrebrist under hjärtstoppet som bland annat påverkar hjärnan, hjärtat och ger en inflammationsreaktion. Idag används riktlinjer från europeiska sällskapet för intensivvård och från det europeiska förbundet för hjärt-lung-räddning för att bedöma den enskilda patientens risk för ett dåligt neurologiskt utfall t.ex. svåra bestående hjärnskador. Bedömningen bygger på neurologiska undersökningar av patienten, röntgenundersökning av hjärnan, olika metoder för att testa hjärnans elektriska impulser och specifika blodprover.

I studie 4 och 5, undersökte vi om ANN kan förutsäga långtidsföljderna efter hjärtstopp. Redan vid inläggningstidpunkten kunde vi se att neurala nätverk förbättrade riskbedömningen jämfört med tidigare utvecklade modeller (studie 4). För de patienter som inte vaknade de första tre dagarna efter hjärtstoppet gav kombinationen av neurala nätverk och blodprover i form av specifika hjärnskademarkörer lovande resultat (studie 5). Även om vi inte hade tillgång till röntgenundersökningar eller undersökningar av hjärnans elektriska impulser, så var resultaten så bra att vissa av våra modeller var jämförbara med nuvarande metoder.

Ett problem med ANN-modeller är att de är mycket komplicerade. I den femte studien använde vi därför en metod som kallas Shapley additive explanations för att förklara orsaken till modellens bedömningar. På så vis fick vi en bättre inblick i hur ANN fungerar. Sammanfattningsvis har vår forskning med neurala nätverk visat lovande resultat för att kunna förbättra riskbedömningen av IVA-patienter, både generellt och för patienter inlagda efter hjärtstopp.

# Acknowledgements and grants

As a PhD student, I faced many challenges and hurdles. Many people were instrumental either directly or indirectly in assisting and supporting me through thick and thin. Thank you all!

**Attila Frigyesi**, my supervisor, friend and statistical mentor, who gave me this opportunity and introduced me to a whole new world of predictive modelling. I truly enjoyed our enthusiastic discussions and appreciated our shared passion for this research field. I truly appreciate your support and guidance. Without you, this thesis would not have been possible.

**Jonas Björk**, my co-supervisor, for your support and teaching, and for all the great discussions.

This work would not have materialised without **the Swedish Intensive Care Registry**. I would like to recognise their important role and acknowledge the thousands of consultants, registrars and nurses who all have contributed to making this database possible. I would also like to thank **all the TTM-trial investigators and contributors** for making the TTM-trial database possible.

**Ola Björnsson**, my friend and research colleague. Thank you for your unconditional support, endless patience and genuine interest in supervised machine learning and coding. I thank you for the midnight oil you burned in assisting me in the final steps of completing papers IV and V.

**Mikael Bodelsson**, my research mentor and professor at our institution. Thank you for your invaluable guidance, coaching and support. In addition, thank you for giving me the opportunity to teach at the university, which was a great experience for me.

**Niklas Nielsen** and **Tobias Cronberg**, you have my deepest gratitude for all your helpful insights, knowledge and resourceful discussions on cardiac arrest prognostication. Your support has been much appreciated.

**Hans Friberg**, my sincere gratitude for funding my research time back when I needed it most. It was an important stepping stone to get me started.

**Andreas Jakobsson**, I can't thank you enough for your genuine willingness to assist me.

**Patrik Johnsson**, my friend and colleague, without whom I would not have received half the grants I got. Thank you for your endless help and support. I really appreciate it.

**Märta Leffler**, my friend and colleague. Thank you for your kindness, enthusiasm and encouragement.

This acknowledgement would not be complete without mentioning my research colleagues **Jesper Johnsson, Gustav Holmgren, Josef Dankiewicz** and the rest of my co-authors. It has been a great pleasure working with all of you.

A special acknowledgement to the present and past heads of the Department of Intensive and Perioperative Care, Skåne University Hospital in Lund: **Anders Rehn, Carolina Samuelsson** and **Marie Martinsson**, for creating the opportunity for research in the department. To my immediate present and past managers, heads of sections: **Dag Winstedt, Ingrid Östlund** and **David Piros**, thank you for allocating time for my research. Without your support, this thesis would not have been possible.

To all of **my friends and colleagues in the Department of Intensive and Perioperative Care**, thank you for your endless energy in creating a good work environment. Thank you for all the good laughs and for cheering me up on gloomy days.

To all **personnel at our department's medical technology section and pain management unit**, thank you for letting me hijack a desk and turn it into a little office.

**Ann Svensson Gustafsson** and **Jan Karlsson**, for your help with data acquisition in studies I and II.

To my roommates, **Björn Bark, Olof Persson, Jan-Erik Kull**, and **Svajunas Statkevicius**, thanks for all the good talks and your support.

To **my friends**, for all your encouragements and for keeping me sane. I really appreciate it.

A special thanks to my **brother Michael** and his fiancé **Line** for the encouragement and support. Thank you, Michael, for letting me use your cool picture as the cover for this thesis.

My parents, **Niels Erik** and **Hanne**, for your unconditional love and encouragement you have given me all these years, and for the utmost support you have given me these past few months. There are no words that can express my gratitude and love for the both of you.

My daughter, **Emelie**, thank you for constantly taking my focus away. Loved every minute of it. I cannot explain how much it means to me.

**Dhashini**, my life-partner, best friend and mother to our daughter Emelie. Thank you for being you. Thank you for your patience and understanding when my mind wanders off. I am looking ahead to more adventures with you by my side.

*Financial support / Grants*

I gratefully acknowledge the funding given by the Regional Research Support from Region Skåne, Lund University and the Royal Physiographic Society of Lund, which have been crucial to the completion of my PhD.

I would also like to acknowledge the travel grant I received from the Knut and Alice Wallenberg Foundation to visit the Massachusetts Institute of Technology. Due to private and very worldwide circumstances, this visit must be postponed until further notice.



# References

1. Rosengart MR. Critical care medicine: landmarks and legends. *Surg Clin North Am.* 2006;86(6):1305-21.
2. Ristagno G, Weil MH. History of Critical Care Medicine: The Past, the Present and the Future. In: Gullo A, Lumb PD, Besso J, Williams GF, editors. *Intensive and Critical Care Medicine: WFSICCM World Federation of Societies of Intensive and Critical Care Medicine.* Milano: Springer Milan; 2009. p. 3-17.
3. Weil MH, Tang W. From intensive care to critical care medicine: a historical perspective. *Am J Respir Crit Care Med.* 2011;183(11):1451-3.
4. Kelly FE, Fong K, Hirsch N, Nolan JP. Intensive care medicine is 60 years old: the history and future of the intensive care unit. *Clin Med (Lond).* 2014;14(4):376-9.
5. Pincock S, Bjørn Aage Ibsen. *The Lancet.* 2007;370(9598):1538.
6. Reisner-Sénélar L. The birth of intensive care medicine: Björn Ibsen's records. *Intensive Care Med.* 2011;37(7):1084-6.
7. Rhodes A, Ferdinande P, Flaatten H, Guidet B, Metnitz PG, Moreno RP. The variability of critical care bed numbers in Europe. *Intensive Care Med.* 2012;38(10):1647-53.
8. Intensivvårdsregistret S. Årsrapport 2019 2020 [cited 2020 Oct 26]. Available from: [https://www.icuregsw.se/globalassets/arsrapporter/arsrapport\\_2019\\_final.pdf](https://www.icuregsw.se/globalassets/arsrapporter/arsrapport_2019_final.pdf).
9. Bauer J, Brüggmann D, Klingelhöfer D, Maier W, Schwettmann L, Weiss DJ, et al. Access to intensive care in 14 European countries: a spatial analysis of intensive care need and capacity in the light of COVID-19. *Intensive Care Medicine.* 2020.
10. Gellerfors M, Gryth D, Lossius HM, Linde J. [Rapid development in prehospital physician staffed intensive care]. *Läkartidningen.* 2016;113.
11. Bouch DC, Thompson JP. Severity scoring systems in the critically ill. *Continuing Education in Anaesthesia Critical Care & Pain.* 2008;8(5):181-5.
12. Keuning BE, Kaufmann T, Wiersema R, Granholm A, Pettila V, Moller MH, et al. Mortality prediction models in the adult critically ill: A scoping review. *Acta Anaesthesiol Scand.* 2020;64(4):424-42.
13. Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. *Crit Care.* 2010;14(2):207.
14. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, 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 Med.* 1996;22(7):707-10.

15. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. *Crit Care Med.* 1984;12(11):975-7.
16. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA.* 1993;270(24):2957-63.
17. Metnitz PG, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3-- From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. *Intensive Care Med.* 2005;31(10):1336-44.
18. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, 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(10):1345-55.
19. 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(5):1297-310.
20. 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(3):827-35.
21. Engerstrom L, Freter W, Sellgren J, Sjoberg F, Fredrikson M, Walther SM. Mortality Prediction After Cardiac Surgery: Higgins' Intensive Care Unit Admission Score Revisited. *Ann Thorac Surg.* 2020;110(5):1589-94.
22. Straney L, Clements A, Parslow RC, Pearson G, Shann F, Alexander J, et al. Paediatric index of mortality 3: an updated model for predicting mortality in pediatric intensive care\*. *Pediatr Crit Care Med.* 2013;14(7):673-81.
23. Starmark JE, Stalhammar D, Holmgren E. The Reaction Level Scale (RLS85). Manual and guidelines. *Acta Neurochir (Wien).* 1988;91(1-2):12-20.
24. Rydenfelt K, Engerstrom L, Walther S, Sjoberg F, Stromberg U, Samuelsson C. In-hospital vs. 30-day mortality in the critically ill - a 2-year Swedish intensive care cohort analysis. *Acta Anaesthesiol Scand.* 2015;59(7):846-58.
25. Intensivvårdsregistret S. Riskjusteringsmodeller inom svensk intensivvård 2020 [cited 2020 Dec 23]. 16.0:[Available from: [https://www.icuregswe.org/globalassets/riktlinjer/riskjustering\\_16.0.pdf](https://www.icuregswe.org/globalassets/riktlinjer/riskjustering_16.0.pdf)].
26. Engerstrom L, Kramer AA, Nolin T, Sjoberg F, Karlstrom G, Fredrikson M, et al. Comparing Time-Fixed Mortality Prediction Models and Their Effect on ICU Performance Metrics Using the Simplified Acute Physiology Score 3. *Crit Care Med.* 2016;44(11):e1038-e44.
27. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128-38.
28. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* 2016;74:167-76.

29. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. *Crit Care Med.* 2006;34(5):1378-88.
30. Goodwin ML, Harris JE, Hernández A, Gladden LB. Blood lactate measurements and analysis during exercise: a guide for clinicians. *J Diabetes Sci Technol.* 2007;1(4):558-69.
31. Andersen LW, Mackenhauer J, Roberts JC, Berg KM, Cocchi MN, Donnino MW. Etiology and therapeutic approach to elevated lactate levels. *Mayo Clin Proc.* 2013;88(10):1127-40.
32. Jeppesen JB, Mortensen C, Bendtsen F, Møller S. Lactate metabolism in chronic liver disease. *Scand J Clin Lab Invest.* 2013;73(4):293-9.
33. Posma RA, Frøslev T, Jespersen B, van der Horst ICC, Touw DJ, Thomsen RW, et al. Prognostic impact of elevated lactate levels on mortality in critically ill patients with and without preadmission metformin treatment: a Danish registry-based cohort study. *Annals of Intensive Care.* 2020;10(1):36.
34. Villar J, Short JH, Lighthall G. Lactate Predicts Both Short- and Long-Term Mortality in Patients With and Without Sepsis. *Infect Dis (Auckl).* 2019;12:1178633719862776-.
35. Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, et al. Developing a New Definition and Assessing New Clinical Criteria for Septic Shock: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):775-87.
36. Xu RY, Zhu XF, Yang Y, Ye P. High-sensitive cardiac troponin T. *J Geriatr Cardiol.* 2013;10(1):102-9.
37. Morrow DA, Cannon CP, Jesse RL, Newby LK, Ravkilde J, Storrow AB, et al. National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines: Clinical characteristics and utilization of biochemical markers in acute coronary syndromes. *Circulation.* 2007;115(13):e356-75.
38. Lara B, Salinero JJ, Gallo-Salazar C, Areces F, Ruiz-Vicente D, Martinez M, et al. Elevation of Cardiac Troponins After Endurance Running Competitions. *Circulation.* 2019;139(5):709-11.
39. Leckie T, Richardson A, Watkins E, Fitzpatrick D, Galloway R, Grimaldi R, et al. High-sensitivity troponin T in marathon runners, marathon runners with heart disease and collapsed marathon runners. *Scand J Med Sci Sports.* 2019;29(5):663-8.
40. Vasile VC, Chai HS, Abdeldayem D, Afessa B, Jaffe AS. Elevated cardiac troponin T levels in critically ill patients with sepsis. *Am J Med.* 2013;126(12):1114-21.
41. Khera R, Agusala V, Cheeran D, Reddy PP, Link MS. Abstract 17130: Diagnostic and Prognostic Utility of Cardiac Troponin in Post-Cardiac Arrest Care. *Circulation.* 2017;136(suppl\_1):A17130-A.
42. Babuin L, Vasile VC, Rio Perez JA, Alegria JR, Chai HS, Afessa B, et al. Elevated cardiac troponin is an independent risk factor for short- and long-term mortality in medical intensive care unit patients. *Crit Care Med.* 2008;36(3):759-65.



43. Docherty AB, Sim M, Oliveira J, Adlam M, Ostermann M, Walsh TS, et al. Early troponin I in critical illness and its association with hospital mortality: a cohort study. *Critical Care*. 2017;21(1):216.
44. Lim W, Qushmaq I, Cook DJ, Crowther MA, Heels-Ansdell D, Devereaux PJ, et al. Elevated troponin and myocardial infarction in the intensive care unit: a prospective study. *Critical care (London, England)*. 2005;9(6):R636-R44.
45. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. *J Family Med Prim Care*. 2019;8(7):2328-31.
46. Kok JN BE, Kosters WA, van der Putten P, Poel M. Artificial Intelligence: Definition, Trends, Techniques and Cases: *Encyclopedia of Life Support Systems (EOLSS)*; 2013.
47. Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artificial Intelligence in Medicine*. 2020;104:101820.
48. Pereira CR, Pereira DR, Rosa GH, Albuquerque VHC, Weber SAT, Hook C, et al. Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification. *Artificial Intelligence in Medicine*. 2018;87:67-77.
49. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*. 2019;9(1):12495.
50. Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. *Br J Gen Pract*. 2018;68(668):143-4.
51. Shimizu H, Nakayama KI. Artificial intelligence in oncology. *Cancer Sci*. 2020;111(5):1452-60.
52. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. 2015;349(6245):255-60.
53. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*. 2018;24(11):1716-20.
54. Hyun S, Kaewprag P, Cooper C, Hixon B, Moffatt-Bruce S. Exploration of critical care data by using unsupervised machine learning. *Computer Methods and Programs in Biomedicine*. 2020;194:105507.
55. Knox DB, Lanspa MJ, Kuttler KG, Brewer SC, Brown SM. Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome. *Intensive care medicine*. 2015;41(5):814-22.
56. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*. 2002;35(5):352-9.
57. Dasgupta A, Sun YV, Konig IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genet Epidemiol*. 2011;35 Suppl 1:S5-11.

58. Rymarczyk T, Kozłowski E, Klosowski G, Niderla K. Logistic Regression for Machine Learning in Process Tomography. *Sensors (Basel)*. 2019;19(15).
59. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 2019;19(1):281.
60. Gutierrez G. Artificial Intelligence in the Intensive Care Unit. *Critical Care*. 2020;24(1):101.
61. Deasy J, Liò P, Ercole A. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation. *Scientific Reports*. 2020;10(1):22129.
62. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*. 2020;2(4):e179-e91.
63. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*. 2019;139(10):e56-e528.
64. Gräsner J-T, Lefering R, Koster RW, Masterson S, Böttiger BW, Herlitz J, et al. EuReCa ONE—27 Nations, ONE Europe, ONE Registry: A prospective one month analysis of out-of-hospital cardiac arrest outcomes in 27 countries in Europe. *Resuscitation*. 2016;105:188-95.
65. Hjärt-Lungräddningsregistret S. Årsrapport 2019 2020 [cited 2020 Dec 14]. Available from: <https://www.hlr.nu/wp-content/uploads/2020/09/Svenska-HLR-registret-%C3%A5rsrapport-2019-publicerad-2020.pdf>.
66. Adnet F, Triba MN, Borron SW, Lapostolle F, Hubert H, Gueugniaud PY, et al. Cardiopulmonary resuscitation duration and survival in out-of-hospital cardiac arrest patients. *Resuscitation*. 2017;111:74-81.
67. Panchal AR, Bartos JA, Cabanas JG, Donnino MW, Drennan IR, Hirsch KG, et al. Part 3: Adult Basic and Advanced Life Support: 2020 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation*. 2020;142(16\_suppl\_2):S366-S468.
68. Nolan JP, Neumar RW, Adrie C, Aibiki M, Berg RA, Bottiger BW, et al. Post-cardiac arrest syndrome: epidemiology, pathophysiology, treatment, and prognostication. A Scientific Statement from the International Liaison Committee on Resuscitation; the American Heart Association Emergency Cardiovascular Care Committee; the Council on Cardiovascular Surgery and Anesthesia; the Council on Cardiopulmonary, Perioperative, and Critical Care; the Council on Clinical Cardiology; the Council on Stroke. *Resuscitation*. 2008;79(3):350-79.
69. Callaway CW, Donnino MW, Fink EL, Geocadin RG, Golan E, Kern KB, et al. Part 8: Post-Cardiac Arrest Care: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation*. 2015;132(18\_Suppl\_2):S465-82.
70. Jennett B, Bond M. Assessment of outcome after severe brain damage. *Lancet*. 1975;1(7905):480-4.

71. Ajam K, Gold LS, Beck SS, Damon S, Phelps R, Rea TD. Reliability of the Cerebral Performance Category to classify neurological status among survivors of ventricular fibrillation arrest: a cohort study. *Scand J Trauma Resusc Emerg Med.* 2011;19:38.
72. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke.* 1988;19(5):604-7.
73. Pareek N, Kordis P, Beckley-Hoelscher N, Pimenta D, Kocjancic ST, Jazbec A, et al. A practical risk score for early prediction of neurological outcome after out-of-hospital cardiac arrest: MIRACLE2. *European Heart Journal.* 2020;41(47):4508-17.
74. Martinell L, Nielsen N, Herlitz J, Karlsson T, Horn J, Wise MP, et al. Early predictors of poor outcome after out-of-hospital cardiac arrest. *Critical care (London, England).* 2017;21(1):96-.
75. Pareek N, Kordis P, Beckley-Hoelscher N, Pimenta D, Kocjancic ST, Jazbec A, et al. A practical risk score for early prediction of neurological outcome after out-of-hospital cardiac arrest: MIRACLE2. *Eur Heart J.* 2020.
76. Dragancea I, Rundgren M, Englund E, Friberg H, Cronberg T. The influence of induced hypothermia and delayed prognostication on the mode of death after cardiac arrest. *Resuscitation.* 2013;84(3):337-42.
77. Lemiale V, Dumas F, Mongardon N, Giovanetti O, Charpentier J, Chiche J-D, et al. Intensive care unit mortality after cardiac arrest: the relative contribution of shock and brain injury in a large cohort. *Intensive Care Medicine.* 2013;39(11):1972-80.
78. Nolan JP, Cariou A. Post-resuscitation care: ERC-ESICM guidelines 2015. *Intensive Care Med.* 2015;41(12):2204-6.
79. Furbass F, Herta J, Koren J, Westover MB, Hartmann MM, Gruber A, et al. Monitoring burst suppression in critically ill patients: Multi-centric evaluation of a novel method. *Clin Neurophysiol.* 2016;127(4):2038-46.
80. Rothstein TL. SSEP retains its value as predictor of poor outcome following cardiac arrest in the era of therapeutic hypothermia. *Critical Care.* 2019;23(1):327.
81. Moseby-Knappe M, Westhall E, Backman S, Mattsson-Carlgren N, Dragancea I, Lybeck A, et al. Performance of a guideline-recommended algorithm for prognostication of poor neurological outcome after cardiac arrest. *Intensive Care Med.* 2020.
82. Zhou SE, Maciel CB, Ormseth CH, Beekman R, Gilmore EJ, Greer DM. Distinct predictive values of current neuroprognostic guidelines in post-cardiac arrest patients. *Resuscitation.* 2019;139:343-50.
83. Bongiovanni F, Romagnosi F, Barbella G, Di Rocco A, Rossetti AO, Taccone FS, et al. Standardized EEG analysis to reduce the uncertainty of outcome prognostication after cardiac arrest. *Intensive Care Med.* 2020;46(5):963-72.
84. Sandroni C, Grippo A, Nolan JP. ERC-ESICM guidelines for prognostication after cardiac arrest: time for an update. *Intensive Care Medicine.* 2020;46(10):1901-3.
85. Kim MJ, Kim T, Suh GJ, Kwon WY, Kim KS, Jung YS, et al. Association between the simultaneous decrease in the levels of soluble vascular cell adhesion molecule-1 and S100 protein and good neurological outcomes in cardiac arrest survivors. *Clin Exp Emerg Med.* 2018;5(4):211-8.

86. Mattsson N, Zetterberg H, Nielsen N, Blennow K, Dankiewicz J, Friberg H, et al. Serum tau and neurological outcome in cardiac arrest. *Ann Neurol*. 2017;82(5):665-75.
87. Grand J, Kjaergaard J, Nielsen N, Friberg H, Cronberg T, Bro-Jeppesen J, et al. Serum tau fragments as predictors of death or poor neurological outcome after out-of-hospital cardiac arrest. *Biomarkers*. 2019;24(6):584-91.
88. Moseby-Knappe M, Mattsson N, Nielsen N, Zetterberg H, Blennow K, Dankiewicz J, et al. Serum Neurofilament Light Chain for Prognosis of Outcome After Cardiac Arrest. *JAMA Neurol*. 2019;76(1):64-71.
89. Kaneko T, Kasaoka S, Miyauchi T, Fujita M, Oda Y, Tsuruta R, et al. Serum glial fibrillary acidic protein as a predictive biomarker of neurological outcome after cardiac arrest. *Resuscitation*. 2009;80(7):790-4.
90. Ok G, Aydin D, Erbuyun K, Gursoy C, Taneli F, Bilge S, et al. Neurological outcome after cardiac arrest: a prospective study of the predictive ability of prognostic biomarkers neuron-specific enolase, glial fibrillary acidic protein, S-100B, and procalcitonin. *Turk J Med Sci*. 2016;46(5):1459-68.
91. Ebner F, Moseby-Knappe M, Mattsson-Carlgrén N, Lilja G, Dragancea I, Uden J, et al. Serum GFAP and UCH-L1 for the prediction of neurological outcome in comatose cardiac arrest patients. *Resuscitation*. 2020.
92. Düring J, Annborn M, Cronberg T, Dankiewicz J, Devaux Y, Hassager C, et al. Copeptin as a marker of outcome after cardiac arrest: a sub-study of the TTM trial. *Critical care (London, England)*. 2020;24(1):185-.
93. Myhre PL, Tiainen M, Pettilä V, Vaahersalo J, Hagve TA, Kurola J, et al. NT-proBNP in patients with out-of-hospital cardiac arrest: Results from the FINNRESUSCI Study. *Resuscitation*. 2016;104:12-8.
94. Bro-Jeppesen J, Kjaergaard J, Stammet P, Wise MP, Hovdenes J, Åneman A, et al. Predictive value of interleukin-6 in post-cardiac arrest patients treated with targeted temperature management at 33 °C or 36 °C. *Resuscitation*. 2016;98:1-8.
95. Annborn M, Dankiewicz J, Erlinge D, Hertel S, Rundgren M, Smith JG, et al. Procalcitonin after cardiac arrest – An indicator of severity of illness, ischemia-reperfusion injury and outcome. *Resuscitation*. 2013;84(6):782-7.
96. Annborn M, Nilsson F, Dankiewicz J, Rundgren M, Hertel S, Struck J, et al. The Combination of Biomarkers for Prognostication of Long-Term Outcome in Patients Treated with Mild Hypothermia After Out-of-Hospital Cardiac Arrest-A Pilot Study. *Ther Hypothermia Temp Manag*. 2016;6(2):85-90.
97. Gaetani L, Blennow K, Calabresi P, Di Filippo M, Parnetti L, Zetterberg H. Neurofilament light chain as a biomarker in neurological disorders. *Journal of Neurology, Neurosurgery & Psychiatry*. 2019;90(8):870-81.
98. Wihersaari L, Ashton NJ, Reinikainen M, Jakkula P, Pettilä V, Hästbacka J, et al. Neurofilament light as an outcome predictor after cardiac arrest: a post hoc analysis of the COMACARE trial. *Intensive Care Medicine*. 2020.
99. Isgro MA, Bottoni P, Scatena R. Neuron-Specific Enolase as a Biomarker: Biochemical and Clinical Aspects. *Adv Exp Med Biol*. 2015;867:125-43.

100. Mastroianni A, Panella R, Morelli D. Invisible hemolysis in serum samples interferes in NSE measurement. *Tumori Journal*. 2020;106(1):79-81.
101. Stammet P, Collignon O, Hassager C, Wise MP, Hovdenes J, Aneman A, et al. Neuron-Specific Enolase as a Predictor of Death or Poor Neurological Outcome After Out-of-Hospital Cardiac Arrest and Targeted Temperature Management at 33 degrees C and 36 degrees C. *J Am Coll Cardiol*. 2015;65(19):2104-14.
102. Wiberg S, Hassager C, Stammet P, Winther-Jensen M, Thomsen JH, Erlinge D, et al. Single versus Serial Measurements of Neuron-Specific Enolase and Prediction of Poor Neurological Outcome in Persistently Unconscious Patients after Out-Of-Hospital Cardiac Arrest - A TTM-Trial Substudy. *PLoS One*. 2017;12(1):e0168894.
103. Nielsen N, Wetterslev J, Cronberg T, Erlinge D, Gasche Y, Hassager C, et al. Targeted temperature management at 33 degrees C versus 36 degrees C after cardiac arrest. *N Engl J Med*. 2013;369(23):2197-206.
104. Nielsen N, Wetterslev J, al-Subaie N, Andersson B, Bro-Jeppesen J, Bishop G, et al. Target Temperature Management after out-of-hospital cardiac arrest--a randomized, parallel-group, assessor-blinded clinical trial--rationale and design. *Am Heart J*. 2012;163(4):541-8.
105. Safari S, Baratloo A, Elfil M, Negida A. Evidence Based Emergency Medicine; Part 5 Receiver Operating Curve and Area under the Curve. *Emergency (Tehran, Iran)*. 2016;4(2):111-3.
106. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. 2019;17(1):230.
107. Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G. Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PloS one*. 2011;6(2):e16110-e.
108. Siegel T, Adamski J, Nowakowski P, Onichimowski D, Weigl W. Prospective assessment of standardized mortality ratio (SMR) as a measure of quality of care in intensive care unit--a single-centre study. *Anaesthesiol Intensive Ther*. 2015;47(4):328-32.
109. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. *Biom J*. 2008;50(4):457-79.
110. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768-77.
111. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-60.
112. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-45.
113. Molnar C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable* 2019.
114. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing,; 2020.

115. Python Core Team. Python: A dynamic, open source programming language. Python version 3.7 ed: Python Software Foundation; 2020.
116. Abadi M, Barham P, Chen JM, Chen ZF, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. Proceedings of Osd'16: 12th Usenix Symposium on Operating Systems Design and Implementation. 2016:265-83.
117. Bartel KYaA. tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights 2020 [R package version 0.12.0:[Available from: <https://CRAN.R-project.org/package=tableone>.
118. Lumley MGaT. forestplot: Advanced Forest Plot Using 'grid' Graphics 2020 [R package version 1.10.1:[Available from: <https://CRAN.R-project.org/package=forestplot>.
119. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
120. López-Ratón M, Rodríguez-Álvarez MX, Suárez CC, Sampedro FG. OptimalCutpoints : An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of statistical software. 2014;61(8):1-36.
121. Keany E. BorutaShap 1.0.14 2020 [Available from: <https://pypi.org/project/BorutaShap/>.
122. Tantau T, editor Graph Drawing in TikZ2013; Berlin, Heidelberg: Springer Berlin Heidelberg.
123. Vallabhajosyula S, Sakhuja A, Geske JB, Kumar M, Poterucha JT, Kashyap R, et al. Role of Admission Troponin-T and Serial Troponin-T Testing in Predicting Outcomes in Severe Sepsis and Septic Shock. J Am Heart Assoc. 2017;6(9).
124. Røsjø H, Varpula M, Hagve T-A, Karlsson S, Ruokonen E, Pettilä V, et al. Circulating high sensitivity troponin T in severe sepsis and septic shock: distribution, associated factors, and relation to outcome. Intensive care medicine. 2011;37(1):77-85.
125. Lengquist M, Lundberg OHM, Spangfors M, Annborn M, Levin H, Friberg H, et al. Sepsis is underreported in Swedish intensive care units: A retrospective observational multicentre study. Acta Anaesthesiol Scand. 2020;64(8):1167-76.
126. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.
127. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature Communications. 2020;11(1):3852.
128. Alves T, Laender AHF, Veloso A, Ziviani N. Dynamic Prediction of ICU Mortality Risk Using Domain Adaptation. 2018 IEEE International Conference on Big Data (Big Data). 2018:1328-36.

