Linköping University Medical Dissertations No. 1637

The significance of risk adjustment for the assessment of results in intensive care.

An analysis of risk adjustment models
used in Swedish intensive care.

Lars Engerström

Department of Medical and Health Sciences
Linköping University, Sweden
Linköping 2018

©Lars Engerström, 2018

Cover illustration: Lars Engerström

Published articles have been reprinted with the permission of the copyright holders.

Printed in Sweden by LiU-Tryck, Linköping, Sweden, 2018

ISBN 978-91-7685-228-6

ISSN 0345-0082

To Carina, Isabelle and Erik

CONTENTS

ABSTRACT	1
SVENSK SAMMANFATTNING	2
LIST OF PAPERS	3
ABBREVIATIONS.....	4
INTRODUCTION.....	5
BACKGROUND.....	7
Risk adjustment.....	7
Intensive care in Sweden	8
History of intensive care risk adjustment	10
Mortality endpoints.....	11
Data sampling window.....	12
Current models.....	13
<i>General intensive care.....</i>	<i>13</i>
<i>Cardiothoracic intensive care.....</i>	<i>16</i>
Risk adjustment in Swedish intensive care	19
<i>Risk adjustment models used in Swedish general intensive care of adults</i>	<i>19</i>
<i>Risk adjustment models used in Swedish cardiothoracic intensive care of adults</i>	<i>19</i>
<i>Risk adjustment models used in paediatric intensive care in Sweden</i>	<i>20</i>
AIMS.....	21
MATERIAL AND METHODS.....	23
The Swedish Intensive Care Registry	23
<i>Quality of SIR data</i>	<i>23</i>
The Swedish National Patient Register	27
Linking of data.....	27
Patients	27
Measures of model performance	28
<i>Discrimination.....</i>	<i>28</i>
<i>Calibration</i>	<i>30</i>
<i>Overall performance measures, accuracy</i>	<i>33</i>

<i>Confidence intervals</i>	34
Missing data	37
<i>Importance of missing data</i>	37
<i>Patterns of missing data</i>	37
<i>Imputation</i>	37
Model validation.....	38
Recalibration of models	38
Model development.....	39
<i>Regression</i>	39
<i>Overfitting</i>	42
<i>Penalisation</i>	42
Methods specific to each study	42
RESULTS	45
The performance of SAPS3 in Swedish intensive care.....	45
The effect of missing data on performance of the model.....	46
The relationship between 30-day and in-hospital mortality.	51
The use of a time fixed mortality endpoint instead of in-hospital mortality for risk adjustment.....	51
The usefulness of the SAPS3 outcome prediction model using different time-fixed outcomes.....	53
Effect of the outcome measure used when comparing intensive care units.	55
Risk prediction after cardiovascular surgery.	57
DISCUSSION	61
The performance of SAPS3 in Swedish intensive care.....	61
The effect of missing data on performance of the model.....	65
The use of a time-fixed mortality endpoint versus in-hospital mortality.	66
Effect of the outcome measure used when comparing intensive care units.	67
Risk prediction after cardiovascular surgery.	68
Quality of SIR data	72
So, what is the significance of risk adjustment in Swedish intensive care?	74
CONCLUSIONS	79
ACKNOWLEDGEMENTS	81
REFERENCES	82

ABSTRACT

To study the development of mortality in intensive care over time or compare different departments, you need some kind of risk adjustment to make analysis meaningful since patient survival varies with severity of the disease. With the aid of a risk adjustment model, expected mortality can be calculated. The actual mortality rate observed can then be compared to the expected mortality rate, giving a risk-adjusted mortality.

In-hospital mortality is commonly used when calculating risk-adjusted mortality following intensive care, but in-hospital mortality is affected by the duration of care and transfer between units. Time-fixed measurements such as 30-day mortality are less affected by this and are a more objective measure, but the intensive care models that are available are not adapted for this measure. Furthermore, how length of follow-up affects risk adjusted mortality has not been studied. The degree and pattern of loss of physiological data that exists and how this affects performance of the model has not been properly studied. General intensive care models perform poorly for cardiothoracic intensive care where admission is often planned, where cardiovascular physiology is more affected by extra corporeal circulation and where the reasons for admission are usually not the same.

The model used in Sweden for adult general intensive care patients is the Simplified Acute Physiology Score 3 (SAPS3). SAPS3 recalibrations were made for in-hospital mortality and 30-, 90- and 180-day mortality. Missing data were simulated, and the resulting performance compared to performance in datasets with originally missing data.

We conclude that SAPS3 works equally well using 30-day mortality as in-hospital mortality.

The performance with both 90- and 180-day mortality as outcome was also good. It was found that the model was stable when validated in other patients than it was recalibrated with.

We conclude that the amount of data missing in the SIR has a limited effect on model performance, probably because of active data selection based on the patient's status and reason for admission.

A model for cardiothoracic intensive care based on variables available on arrival at Swedish cardiothoracic intensive care units was developed and found to perform well.

SVENSK SAMMANFATTNING

Om man ska studera utvecklingen av mortaliteten i intensivvården över tid eller jämföra olika avdelningar behöver man någon form av riskjustering för att analysen ska bli meningsfull eftersom patienternas överlevnad varierar med sjukdomens prognos. Med hjälp av riskjusteringsmodeller kan en förväntad mortalitet beräknas. Den faktiskt observerade mortaliteten kan sedan jämföras med den förväntade mortaliteten, vilket ger en så kallad riskjusterad mortalitet.

Internationellt används vanligtvis sjukhusmortalitet för uppföljning av vård. Sjukhusmortaliteten påverkas av vårdtid och överflyttning mellan enheter. Tidsfixerade mått som 30-dagarsmortalitet påverkas mindre av sådant och är ett mer objektiva mått, men de intensivvårdsmodeller som finns är inte anpassade för detta mått. Hur längden på uppföljningstiden påverkar den riskjusterade mortaliteten är inte heller studerat. Hur stort bortfall av fysiologiska data som finns, vilka mönster det har och hur det påverkar modellen är inte ordentligt studerat. De generella intensivvårdsmodellerna presterar sämre för thoraxintensivvård.

Den modell som vanligen används i Sverige för allmän intensivvård av vuxna patienter är Simplified Acute Physiology Score 3 (SAPS3). SAPS3 kalibrerades om för sjukhusmortalitet, 30-, 90- och 180-dagarsmortalitet.

Vi jämförde prestanda hos SAPS3 för att beräkna förväntad 30-dagarsmortalitet med prestanda för att beräkna förväntad sjukhusmortalitet som modellen är utvecklad för. Detta visade att prestanda hos SAPS3 är lika bra med 30-dagarsmortalitet som utfall som med sjukhusmortalitet. Vi undersökte hur SAPS3 fungerar med 90-dagarsmortalitet liksom även 180-dagarsmortalitet som utfall och kunde konstatera att SAPS3 har likartade prestanda för beräkning av förväntad 30-, 90- och 180-dagarsmortalitet. Vi kontrollerade också hur bra prestanda SAPS3 har i ett annat patientmaterial än det som användes när modellen kalibrerades om och modellen visade sig vara stabil.

Vi undersökte hur saknade fysiologiska variabler påverkar SAPS3 modellens prestanda. Slutsatsen var att påverkan av saknade variabler i den mängd de förekommer i Sverige är begränsad, troligen för att det utifrån patientens status och inläggningsorsak sker ett urval av om man behöver ta vissa prover.

Vi utvecklade en modell med bra prestanda för thoraxintensivvård baserad på variabler som finns tillgängliga vid ankomst till thoraxintensivvårdsavdelningen.

LIST OF PAPERS

- I. Rydenfelt K, Engerström L, Walther S, et al: In-hospital vs. 30-day mortality in the critically ill - a 2-year Swedish intensive care cohort analysis. *Acta Anaesthesiol Scand* 2015; 59:846–858
- II. Engerström L, Kramer AA, Nolin T, 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:e1038–e1044
- III. Engerström L, Nolin T, Mårdh CL, et al: Impact of Missing Physiologic Data on Performance of the Simplified Acute Physiology Score 3 Risk-Prediction Model. *Crit Care Med* 2017; 45:2006–2013
- IV. Engerström L, Freter W, Sellgren J, Sjöberg F, Fredrikson M, Walther S: Development and validation of a model for the prediction of 30-day mortality on admission to the intensive care unit after cardiac and cardiovascular surgery. (submitted to *Br J Anaesth*, 2018)

ABBREVIATIONS

aROC	Area under the Receiver Operating Characteristics curve
ASA	American Society of Anesthesiologists
BE	Base Excess
BSA	Body Surface Area
CABG	Coronary Artery Bypass Grafting
CVP	Central Venous Pressure
ECC	ExtraCorporeal Circulation
ECMO	ExtraCorporeal Membrane Oxygenation
EMR	Estimated Mortality Rate
ESICM	European Society of Intensive Care Medicine
GCS	Glasgow Coma Scale
HR	Heart Rate
IABP	Intra-Aortic Balloon Pump
ICNARC	Intensive Care National Audit and Research Centre
ICU	Intensive Care Unit
IQR	InterQuartile Range
MAP	Mean Arterial Pressure
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
MPM	Mortality Prediction Model
O ₂ EF	Oxygen Extraction Fraction
OMR	Observed Mortality Rate
RLS85	Reaction Level Scale 85
SFAI	Swedish Association for Anaesthesia and Intensive care
SIR	The Swedish Intensive Care Registry
SMR	Standardised Mortality Ratio
SPAR	Swedish National Population Register
VLAD	Variable Life Adjusted Display
WBC	White Blood Cell

INTRODUCTION

In the early days of medicine, the effect of medical care was seldom measured sometimes resulting in a worse outcome than the disease itself. One of the more famous examples is Ignaz Semmelweis' observation in the mid-1800s that the mortality rate during childbirth in one obstetric clinic was more than double the rate in another clinic (16% versus 7%). The cause of death was mostly puerperal fever. He also noted that the obstetricians and medical students at the high mortality clinic often went directly to the delivery suite after performing an autopsy and had an odour on their hands despite handwashing with soap and water. He therefore hypothesised that "cadaverous particles" were transmitted via the hands of doctors and students from the autopsy room to the delivery theatre causing the puerperal fever. Consequently, Semmelweis recommended that hands should be scrubbed in a chlorinated lime solution before every patient contact. After the implementation of this measure, the mortality rate fell to 3% in the clinic most affected and remained low thereafter [1].

One of the earliest advocates of analysing outcome data was Florence Nightingale (1820-1910) [2]. She noted a difference in mortality rate between hospitals, with lower mortality rates in the smaller county hospitals compared to the larger hospitals in London. She made the important observation that crude mortality is not an accurate reflection of outcome and suggested that not only patient outcome but also the severity of the disease should be measured.

It was later found that it is crucial to measure outcome. The outcome measured may differ according to type of medical care. The most obvious outcome is mortality. But even mortality rates could be measured at different points in time or at discharge from ward or hospital. Mortality is not always relevant. For example, mortality may not be relevant as an outcome measure in cases where death is infrequent. In other cases, complication rates are more relevant when measuring the quality of medical care. One problem with measuring complication rates is that careful recording of complications does not result in reward but rather punishment in the form of poor rating compared to others. Other relevant outcome measures could be quality of life, strength or ability to carry out daily activities. In cancer surgery it could be relapse, in cardiac care it could be new infarction. Compliance with guidelines also indicates good medical care, but the ultimate measure is outcome.

The measurement of outcome is still the basis for results. Mortality rates can be compared when disease groups are homogenous, but the in-

tensive care population is very heterogeneous. Intensive care is a form of medical care where patients have very varying degrees of illness and causes of admission. For example, there is sepsis with high mortality and intoxication with low mortality, and the case-mix differs between different units and at different times. Comparing raw mortality rates between units is therefore not very valuable in intensive care.

BACKGROUND

Risk adjustment

Risk adjustment models were developed to take the degree of disease into account when comparing mortality rates. If one is to study the development of quality in the medical care of patients with varying degrees of illness over time or compare different departments, some form of risk adjustment is required to make any analysis meaningful. Mortality without risk adjustment is of limited value as a measure of quality of care since patient survival varies with the degree of illness. Risk adjustment, also called risk prediction or outcome prediction, enables us to calculate the predicted mortality rate for a particular patient group and then compare the observed mortality rate of that group with the predicted mortality rate for the group. The predicted mortality rate is calculated from known patient parameters such as age, chronic morbidity and a description of the acute disease such as reason for admission and physiological parameters.

The predicted mortality rate can be used to describe how ill patients are on an intensive care unit (ICU). The predicted mortality can be compared with the actual mortality observed and used to calculate the standardised mortality ratio (SMR). SMR is calculated as the ratio between observed and expected (predicted) mortality. Expected and observed mortality can also be used to describe outcomes graphically with variable life-adjusted display (VLAD, Figure 1).

Beyond benchmarking, ICU risk prediction models have a role in risk adjustment and risk stratification in randomised controlled trials, and when adjusting for confounders in non-randomised, observational research. Risk adjustment models may have a role in communicating risk, but their suitability for individual patient decision-making is limited [3].

Figure 1. Example of variable life-adjusted display (VLAD) of different intensive care units (ICUs). Each color represents a unit. A line moving upwards indicates more survivors than expected.

Risk adjustment models cannot accurately predict the outcome of a single patient. The task of the risk adjustment model is to accurately calculate the probability of a particular outcome based on the information available.

Intensive care in Sweden

In Sweden, intensive care has been defined by the Swedish Association for Anaesthesia and Intensive Care (SFAI) as care that is aimed to prevent and treat failure in one or more organ systems so that continued life can be meaningful from the patient's point of view [4]. Intensive care should be available 24 hours a day, all year, with retained quality. The care undertaken at ICUs is to be evaluated, quality-assured and quality-reviewed. Data collection is a prerequisite for the department's developmental work, and routine follow-up of health and quality data should therefore be performed. All ICUs should be connected to an appropriate quality register and, aim to deliver high quality data.

Sweden has 84 ICUs [5], 80 of which are affiliated to the Swedish Intensive Care Registry (SIR) and regularly send data. All 65 general ICUs, 5/8 cardiothoracic ICUs, 4/4 paediatric ICUs, 4/5 neuro ICUs and 2/2 burn ICUs are affiliated members. The SIR-affiliated units had a total of 516 beds on one typical weekday in 2017. In November 2017, Sweden had 10 112 669 inhabitants making 5.1 intensive care beds per 100 000 inhabitants [6]. This is low compared to figures reported from other European countries [7]. It is also lower than that reported in a questionnaire-study in Sweden 2001-2002 when there were 7.2 ICU beds per 100 000 inhabitants. The general ICUs have 409 beds, the affiliated cardiothoracic ICUs have 44 beds, the paediatric ICUs have 29 beds, the affiliated neuro ICUs have 28 beds, and the burn ICUs have 6 beds.

Compared to international intensive care, the length of stay on Swedish intensive care units is very short [8].

Postoperative care less than 24 hours is not classed as intensive care in Sweden if it does not include respiratory care, continuous renal replacement therapy, or more than 6 hours of inotropic or vasoactive support.

Figure 2. Intensive care units in Sweden.

History of intensive care risk adjustment

In 1941 the American Society of Anesthesiologists (ASA) published the first version of a 'physical status' classification for patients about to undergo surgery. The same year, Meyer Saklad described the ASA PS grading of a patient's physical state in *Anesthesiology* [9].

In 1952, paediatric anaesthetist Virginia Apgar devised the well-known Apgar-score to better predict outcome of new-born infants [10].

The concept of intensive care arose from the Copenhagen polio epidemic of 1952, which resulted in hundreds of victims developing respiratory failure. Over 300 patients required artificial ventilation for several weeks provided by 1,000 students employed to hand ventilate via a tracheostomy. By 1953, Bjorn Ibsen, the anaesthetist who suggested that positive pressure ventilation should be the treatment of choice during the epidemic, set up the first ICU in Europe [11].

The current ASA PS classification was proposed by Dripps et al in 1961 and adopted by the ASA in 1962 [9,12].

Many ICUs were established in the United States during the 1960s. Because of lack of adjustment for severity of illness, it was difficult to demonstrate improvement in survival due to intensive care [13].

To account for this, work began in 1977 to develop a severity of illness score. This resulted in the first widely used intensive care risk adjustment model called the Acute Physiology And Chronic Health Evaluation (APACHE), published in 1981 [14]. Apache I was developed and validated in a cohort of 805 patients on two ICUs.

During the 1980s the APACHE I system needed refinement and external validation. APACHE II was published in 1985 using data from 5815 ICU admissions at 13 hospitals [15]. The number of physiological measures of severity was reduced from 34 to 12 and mortality prediction was adjusted for 44 diagnoses.

The need to simplify severity measurement and mortality prediction was also emphasised by the development of the Simplified Acute Physiology Score (SAPS) [16] in 1984 and the Mortality Prediction Model (MPM) [17] in 1985. Following APACHE II's publication, independent investigators reported important shortcomings: lack of adjustment for patient selection, location before ICU admission, lead-time bias, and concerns about the timing of data collection [18].

APACHE III was published in 1991 and achieved greater prognostic accuracy than its predecessor at the expense of increased model complexity [18]. Refinements included assessment of the predictive impact of measurement timing, missing data, non-linear weighting (splines) of

physiological variables, and expanded adjustment for ICU admission diagnosis.

SAPS II and MPM II were developed using international data and published 1993 [19,20]. They did not use extensive information about diagnosis and emphasised the need to limit complexity. Opinions diverged about whether prognostic models should be simple or complex, and data collection manual or automated.

Changes in clinical practice made the APACHE III equations poorly calibrated. Instead of simple recalibration, APACHE IV was developed, with expanded ICU admission diagnostic groups, added and refined predictor variables. APACHE IV was published in 2006 [21].

Project IMPACT, a database aimed at describing and measuring the care of ICU patients, was developed by the Society of Critical Care Medicine in the U.S. in 1996 [22], and the data was used for the development of MPM III which was published in 2007 [23].

In 2005, European researchers developed SAPS 3 using international data [24]. These contemporary models continued to differ; APACHE IV remained complex, whereas MPM III and SAPS 3 emphasised simplicity.

In 2007 ICNARC in UK published their own risk adjusted model using raw data from APACHE II, APACHE III, SAPS 2 and MPM II for use in the Case-Mix Programme [25]. A new ICNARC model was published in 2015 as the ICNARC_{CH-2014} model,[26] and has subsequently been recalibrated as the ICNARC_{CH-2015} model using data for admissions to ICUs between 1st April 2013 and 31st March 2014 .

Mortality endpoints

Which mortality endpoint is the most relevant? Traditionally, in-hospital mortality has been used in most models since in many settings it is difficult to follow patients once they have left hospital. More recent work has shown that in-hospital mortality is biased by hospital discharge policies. For example, a relatively high percentage of patients discharged to another hospital will result in a lower observed in-hospital mortality rate [27,28].

There is thus good reason to investigate outcome endpoints that are fixed over time, for example 30, 90 or 180 days after ICU admission, as they are less affected by administrative factors such as hospitalisation [29].

Data sampling window

Allowing longer periods for sampling of physiological data (24 hours) may reduce the number of missing values, but there are problems described with long sampling times (Boyd and Ground, lead-time bias) [30,31]. The Boyd and Ground effect describes how poor medical care during the first 24 hours on the ICU may result in more abnormal physiological parameters. If these physiological parameters are entered into the risk adjustment system during this period, the severity score may increase due to bad care. A higher score will give an increased estimated risk which would give the impression of better quality of care.

Lead-time bias describes a problem when treatment and measurement are started at different times. If treatment is started before admission to ICU and measurement begins on arrival at the ICU, appropriate treatment before ICU will result in better physiological parameters, a lower estimated mortality risk and therefore an impression of poorer care.

The APACHE and ICNARC models use 24 hours for data sampling while the SAPS3 model uses two hours, from one hour before ICU admission to one hour after ICU admission [21,24,32].

Current models

General intensive care

Since case-mix, therapies and so on, successively change over time, model performance decreases and after a while updating is required if the model is to continue to perform well. The most updated versions of the common models for risk adjustment in general intensive care today are Apache IV (2006), SAPS3 (2005), MPM III (2007) and ICNARC (2015). Age limits and variables used in these models are shown in Table 1.

APACHE IV

APACHE IV was developed using a total of 131,618 consecutive ICU admissions to 104 ICUs in 45 U.S. hospitals during 2002 and 2003, of which 110,558 met inclusion criteria and had complete data. ICU Day 1 information and multivariable logistic regression was used to estimate the probability of in-hospital death for randomly selected patients who comprised 60% of the database (development dataset).

The accuracy of APACHE IV predictions was assessed by comparing observed and predicted in-hospital mortality for the excluded patients (validation set). The area under the receiver operating characteristic curve (aROC) was 0.88 and (Hosmer-Lemeshow C statistic = 16.9, $p=0.08$ in the validation dataset. The paper was published in 2006 [21].

Mortality Prediction Model (MPM) III

Was developed using data from 124,855 patients admitted to 135 ICUs at 98 hospitals between 2001 and 2004. All but four hospitals were in the United States; three were Canadian and one Brazilian. The outcome was in-hospital mortality.

In the validation dataset, the aROC was 0.823 (95% confidence interval, 0.818–0.828), the Hosmer-Lemeshow statistic was 11.62 ($p=0.31$), and the SMR 1.018 with a 95% confidence interval of 0.996 –1.040. The paper was published in 2007 [23].

ICNARC model

The new ICNARC model was published in 2015 as the ICNARCS_{H-2014} model. A total of 155,239 admissions to 232 adult ICUs in England, Wales and Northern Ireland between January and December 2012 were used to develop a risk model. The model was validated using 90,017 admissions between January and September 2013. Data were recorded during the first 24 hours on the ICU.

The final model incorporated 15 physiological predictors modelled with continuous non-linear models.

Readmissions of the same patient during the same acute hospital stay and admissions missing ultimate acute hospital outcome were excluded from comparisons of observed and expected mortality. aROC was 0.885 and Brier's score 0.108 [32].

The model has later been recalibrated as the ICNARC_{H-2015} model using data from 134,063 admissions to 238 ICUs between 1st April 2013 and 31st March 2014.

SAPS3

SAPS3 was developed in cooperation with the European Society of Intensive Care Medicine (ESICM) between 2002 and 2004. It was developed on data from a total of 16,784 patients consecutively admitted to 303 ICUs in 35 countries from 14th October to 15th December 2002. All patient categories were included, including cardiac surgery which accounted for 8.4% of the total cohort. More than 70% of patients were cared for on European ICUs. Patients younger than 16 years were excluded.

Twenty variables are included in the model. The degree of physiological disturbance is recorded over a 2-hour period in connection with initiation of intensive care (1 hour before and after arrival). aROC was 0.848 and the paper was published in 2005 [24].

Table 1. Age limits and variables included in the general intensive care models.

Model	APACHE IV	MPM III	ICNARC	SAPS3
Age limits	At least 16 years	At least 18 years	At least 16 years	At least 16 years
Patient characteristics	Age	Age	Age, dependency	Age
Acute physiology	HR, MAP, temperature, oxygenation, haematocrit, WBC, RR, creatinine, urine output, BUN, sodium, albumin, bilirubin, glucose, acid base abnormalities, GCS.	HR, SBP, coma/deep stupor	HR, SBP, temperature, RR, oxygenation, pH, PaCO ₂ , lactate, urine output, urea, creatinine, sodium, WBC, platelets, GCS, sedation.	HR, SBP, GCS, bilirubin, temperature, creatinine, WBC, platelets, pH, oxygenation.
Chronic disease	AIDS, cirrhosis, hepatic failure, immunosuppression, lymphoma, leukemia or myeloma, metastatic tumor.	Chronic renal insufficiency, cirrhosis, metastatic neoplasm.	Chronic liver disease, metastatic disease, haematological malignancy.	Cancer, cancer therapy, chronic heart failure, AIDS, hematological cancer, cirrhosis.
Circumstances of the ICU admission	116 ICU admission diagnosis, location before ICU, LOS before ICU, Emergency surgery, thrombolytic therapy for patients with AMI, mechanical ventilation	ARF, cardiac dysrhythmia, CVI, GI bleed, intracranial mass effect, CPR before admission, mechanical ventilation, medical or unscheduled surgical admission, full code	CPR prior to admission, location prior to admission, urgency of admission, primary reason for admission.	Location before ICU, LOS before ICU, vasoactive therapy before ICU. Planned/unplanned, reasons for ICU admission, surgical status, site of surgery, acute infection on ICU admission

WBC = white blood cell count. GCS = Glasgow Coma Scale. HR = heart rate. MAP = mean arterial blood pressure. SBP = systolic blood pressure. RR = respiratory rate. LOS = length of stay. CVI = cerebrovascular incident. ARF = acute renal failure. BUN = blood urea nitrogen. GI = gastrointestinal. CPR = cardio pulmonary resuscitation. PaCO₂ = partial pressure of arterial carbon dioxide. ICU = intensive care unit. Full code = no restrictions on therapies or interventions at the time of ICU admission. AMI = acute myocardial infarction.

Cardiothoracic intensive care

The general intensive care risk adjustment models are usually less applicable to cardiothoracic intensive care which has a different case-mix [33,34]. This can also be explained by the fact that the acute pathophysiological consequences of cardiopulmonary bypass are transient, and many physiologic changes may be masked by multiple system support devices such as intra-aortic balloon pump, extracorporeal membrane oxygenation (ECMO) and ventricular assist devices. For this reason more specific cardiothoracic intensive care risk adjustment models exist. An overview of age limits and included variables in the described models is shown in Table 2.

Higgins Intensive Care Unit Risk Stratification Score

Higgins Intensive Care Unit Risk Stratification Score was developed on a cohort of 4,918 patients who underwent coronary artery surgery, with or without simultaneous valve or carotid surgery, at a cardiovascular centre in the United States between 1st January, 1993, and 31st March, 1995. The model was developed on a dataset from 2440 patients and the model was validated on 2,125 patients at the same hospital. In-hospital mortality was used as the outcome variable, but this was low (only 78 patients died in the development set) which limited the number of variables included. The score was compiled from simple measures with a single cut-off.

The aROC was 0.85 in the validation cohort. The model was published in 1997 [35].

CASUS

The model was based on all consecutive adult patients admitted after cardiac surgery to a cardiothoracic surgical ICU with 13 beds at the University Hospital of Cologne, Germany, over a period of three years. A total of 3,230 patients were admitted. Evaluation of variables was performed using patients in the development set who stayed on the ICU for at least 24 hours. From April 1999 to May 2000 (development set), 1,069 patients were admitted to the ICU. Of these patients, 384 had an ICU stay longer than 24 hours. From May 2000 to May 2001 (validation set I), 1,057 patients were admitted to the ICU. The validation set II (February 2002 to February 2003) consisted of 1,104 patients with a mortality rate of 4.4% (49 patients). The mean length of ICU stay was 3.2 days. There were no missing data in all three sets.

The aROC for validation set II was 0.90 on the operative day. The paper was published in 2005 [36].

ARctIC

The ARctIC model was developed using data from the Case-Mix Programme which collects intensive care data for England, Wales and Northern Ireland. A predictive model for hospital mortality was derived from variables measured in 17,002 patients within 24 hours of admission to five cardiothoracic ICUs between 1st January 2010 and 31st December 2012 [37]. Patients aged less than 16 years and readmissions to the ICU during the same acute hospital stay were excluded.

The final model included 10 variables. Additional interaction terms between creatinine, lactate, platelet count and cardiac surgery as the admitting diagnosis were included. Missing data were imputed and restricted cubic splines used.

The model was validated against 10,238 other admissions to six cardiothoracic ICUs between 1st January 2013 and 30th June 2014, for which the aROC (95% CI) was 0.904 (0.89-0.92) and the Brier score was 0.055, while the slope and intercept of the calibration plot were 0.961 and -0.183, respectively. The paper was published in 2016 [37].

POCAS

A prospective open-cohort study including 920 patients who had undergone cardiac surgery with cardiopulmonary bypass carried out between January 2009 and January 2011, at the Hospital Clínico Universitario, Valladolid (Spain). The predicted outcome was in-hospital mortality (90 days) which was 9%. Four predictors were included in the model. The aROC was 0.890 while validated in the full development cohort. The paper was published in 2013 [38].

A score to estimate 30-day mortality after intensive care admission after cardiac surgery (Lamarche)

Preoperative and intraoperative data from 30,350 patients in four hospitals were used to build a model estimating 30-day mortality after cardiac surgery. 60 % of the patients were used as the development set 40 % as the validation set. Death occurred in 2.6% of patients (n = 790).

When applied after surgery at ICU admission, the model had aROC of 0.86 in both development and validation sets [39].

Table 2. Age limits and variables included in the cardiothoracic intensive care models.

	Higgins	CASUS	ARCTIC	POCAS	Lamarche
Age limits	None	At least 18 years	Not children	At least 18 years	At least 18 years
Patient characteristics	Age, BSA		Age, dependency		Age, sex
Chronic disease / preoperative factors	Number of prior heart operations, previous vascular surgery, creatinine, albumin				PHT, PVD, renal dysfunction, diabetes, peptic ulcer, alcohol abuse, refusal of blood
Intra-operative factors	CPB-time				Complication, inotropes, vasopressor, RBC transfusion, CPB time, surgical procedure
Circumstances of ICU admission			Location prior to admission		Emergency status
Acute physiology	Oxygenation, HR, CI, CVP, bicarbonate	Oxygenation creatinine, bilirubin, PAR, lactate, platelets, neurology, dialysis	MAP, pH, lactate, creatinine, WBC, platelets, GCS score	MAP, bicarbonate, lactate, INR	
On admission	IABP	IABP, VAD			IABP, VAD or ECMO

RBC = red blood cell. CPB = cardiopulmonary bypass. IABP = Intra-aortic balloon pump. MAP = mean arterial pressure. WBC = white blood cell count, CVP = central venous pressure. PAR = $HR \times CVP / MAP$. VAD = ventricular assist device. PHT = pulmonary hypertension. ECMO = extracorporeal membrane oxygenation. HR = heart rate. INR = the International Normalised Ratio. CI = cardiac index. PVD = peripheral vascular disease.

Risk adjustment in Swedish intensive care

The models commonly used in Sweden are: for general intensive care adult patients (≥ 16 years) the SAPS3; and for adult patients (≥ 16 years) who are admitted to cardiac surgery using extracorporeal circulation (ECC), a modified Higgins Intensive Care Unit Risk Stratification Score (Higgins-ICU).

Risk adjustment models used in Swedish general intensive care of adults

In Swedish intensive care, the APACHE II Risk Adjustment Model has been used since at least the late 1980s when the first efforts to create a national coherent register were made. APACHE II was modified for Swedish conditions, for example the GCS was supplemented with RLS85 for assessment of awareness. When the model's function was analysed in 2005 and 2011, discrimination was good (aROC 0.81) but calibration was unsatisfactory. For example, risk was overestimated at the higher age intervals. In 2008 the SIR began to use the SAPS3. Since 2012, the SIR has switched to using only the SAPS3 model for risk adjustment of general intensive care of patients ≥ 16 years.

Risk adjustment models used in Swedish cardiothoracic intensive care of adults

A significant proportion of cardiovascular patients were included in the cohort from which SAPS3 was developed, but SAPS3 has still proved to be poor in this patient category [33]. The SIR therefore uses a modified Higgins-ICU for risk adjustment during the course of care after cardiac surgery. In this context cardiac surgery means cardiac surgery using ECC.

In 2002, Sellgren and co-workers in Gothenburg adapted the Higgins-ICU for Swedish conditions (unpublished, Table 13), from there the model spread to other Swedish cardiothoracic ICUs. Because discrimination was poor (aROC 0.74 for a joint Swedish cardiothoracic intensive care cohort in 2009), complete recalibration was made by Freter and co-workers where breakpoints and scoring were modified.

This model was developed from 8,210 ICU admissions to four cardiothoracic ICUs during 2007-2010. Compared to the original model, there are three important changes. Missing data are registered as normal, the outcome is survival 30 days after arrival to the ICU instead of in-hospital survival and types of cardiac surgery other than coronary surgery are included in the model. The recalibrated Higgins-ICU model functioned well with an aROC of 0.86, a Cox's calibration regression slope 0.97 and an intercept of -0.11 when evaluated in random thirds of the development dataset (unpublished).

However, it is likely that further adjustment of Higgins-ICU will be needed as the model is based on a small number of outcomes (196 deaths, though this is considerably more than in the original Higgins-ICU development set). The risk of a limited number of outcomes is that coefficients in the model become too optimistic, *i.e.* it is overfitted (see Methods) for the current database. This is seen as a flatter Cox' slope when checking the calibration on recent data. This can be adjusted either by shrinkage of the coefficients in the development dataset or by simply making an overall first-degree recalibration using recent data.

Risk adjustment models used in paediatric intensive care in Sweden

The most widely used models for risk assessment in adult intensive care are developed from patient cohorts that rarely include younger patients. Although SAPS3 included patients aged 16 years and over, only 2 % of patients were 16-20 years old. The cohort from which the Higgins-ICU was developed completely lacks younger patients.

Risk adjustment in paediatric intensive care is not evaluated in this thesis. PIM2 was used in the SIR for all intensive care of children <16 years until 2015 [40]. In 2016 PIM2 was replaced by PIM3 [41]. Mortality in paediatric intensive care is low, requiring many admissions to achieve decent confidence intervals for satisfactory evaluation and recalibration. A PIM4 may be on its way.

AIMS

The overall aim of this thesis was to see how well risk adjustment systems work in Swedish intensive care and examine their importance for the assessment of intensive care results. The following specific questions were addressed:

- How well does SAPS3 perform in Swedish intensive care – how good is its discrimination?
- Is SAPS3 well calibrated?
- Is a Swedish recalibration of SAPS3 needed to improve reliability?
- How extensive is the loss of physiological variables in SAPS3, what is the pattern and what reasons lie behind the loss? How is the performance of the model affected by the loss?
- What is the relationship between 30-day and in-hospital mortality?
- Can a time-fixed mortality endpoint (for example 30-day mortality) be used instead of in-hospital mortality for risk adjustment?
- Is the SAPS3 outcome prediction model useful for benchmarking purposes using different time-fixed outcomes such as 30 days, 90 days and 180 days after admission to the ICU?
- How are comparisons between intensive care units affected by the outcome measure used?
- Can we develop and validate a 30-day mortality risk prediction model with good performance that can be applied to patients undergoing all types of cardiovascular surgery using extracorporeal circulation (ECC) based on data we have already registered?

MATERIAL AND METHODS

The source of data in this thesis was the Swedish Intensive Care Registry's (SIR) [42] database. From this database, which includes more than 200,000 entries from 2009-2016, patients were grouped according to the risk adjustment model used. All raw data coupled to the risk adjustment model used as well as information linked to the date of intensive care, including outcomes, were extracted and analysed.

The Swedish Intensive Care Registry

The SIR is a national database that prospectively collects data from individual intensive care admission records. Data sent to SIR is collected at the bedside using local registration systems during each care session. Transfer of information is encrypted via a specified protocol that also includes validation of data at three different levels. Validation initially initiates correction of primary data which are then re-exported to SIR and inserted into the master database. Follow-up of mortality is done on a weekly basis via the Swedish National Population Register (SPAR) and data from follow-up are continuously added to the register with no limitation in follow-up time. SIR operates within the legal framework of Swedish National Quality Registers [43].

Quality of SIR data

To achieve good quality, a register must cover a large proportion of the patients it is expected to represent, and the data registered must reflect what it is meant to measure. There may be missing predictor data, incorrect predictor data and missing follow-up data. Regarding the predictors used in this thesis, in SAPS3 there are physiological data which could be missing, registered even if sampled outside the two-hour sampling window or simply incorrect. SAPS3 also depends on reason for admission which may be deemed differently from physician to physician. There are many factors to consider when evaluating the quality of a clinical database. A framework for analysing the quality of clinical databases was published in 2002 [44].

The population in the register must be representative of the country. This is fulfilled since all general ICUs, all burn ICUs, all paediatric ICUs, 80 % of the neuro ICUs and 88 % of the cardiothoracic ICUs are members.

The completeness of SIR is good, units participating in SIR recruit consecutive admissions and very few patients refuse registration of their data.

The variables should preferably include identifier, administration information, condition, intervention, short-term outcome, major known confounders, long-term outcome; SIR covers these important variables.

Registered data should be as complete as possible. As described above, there are data missing in SIR regarding physiological parameters in the SAPS3 model and the cardiothoracic intensive care model. The degree and effect of this loss of data will be described later. Administrative data and reason for admission are mandatory and complete.

Data should preferably be recorded in raw format and not categorised. SIR collects continuous data in raw format.

The variables should be explicitly defined. SIR has guidelines with definitions of all variables used.

There should be explicit rules defining how the variables are recorded, for example timing of physiological measurements. This is described in the guidelines presented by SIR.

Coding of conditions and interventions should be reliable. This is a major concern regarding the SIR database since reliability has not been evaluated by external validation.

Observation of primary outcomes should be independent. This is fulfilled in SIR since all deaths are automatically imported from the SPAR.

The data should be extensively validated. SIR data are validated as far as being within specified limits and not illogical. External validation of coding systems for assessment of reason for admission to the ICU and for comparison of measurements recorded in SIR including laboratory results and electronic healthcare records, however, has not been carried out.

Table 3, Performance of the SIR database.

level	1	2	3	4
A. Representative of country				
B. Completeness of recruitment				
C. Variables included				
D. Completeness of variables				
E. Collection of raw data				
F. Explicit definitions				
G. Explicit rules				
H. Reliability of coding				
I. Independence of observations				
J. Data validation				

My assessment of the performance of the SIR database during the period studied against the Directory of Clinical Databases criteria [44].

SIR coverage

Present coverage of Swedish ICUs in SIR is good. At the beginning of this work, coverage was somewhat lower, but increased during the study period (Figure 3). The ICUs that joined later were mostly smaller ones.

Study I; by 2010, SIR had similar coverage for general intensive care in local (89%), county (96%) and university hospitals (100%).

When Studies II and III were carried out, 63 of 78 ICUs (80.7%) having general intensive care patients were included.

During the period of Study IV, 5 of 8 cardiothoracic ICUs sent data to the SIR, but only three sent data continuously.

At present (2018) Sweden has 84 ICUs [5], 80 (95%) of which are affiliated to the SIR and regularly send data. All 65 general ICUs, 5 of 8 cardiothoracic ICUs, 4 of 4 paediatric ICUs, 4 of 5 neuro ICUs and 2 of 2 burn ICUs are members.

Figure 3. SIR coverage of intensive care units (ICUs) over time. Intensive care units registering general intensive care or cardiothoracic intensive care admissions are included. The ICUs included are general, cardiothoracic, burn, paediatric, and neuro.

Missing data and missing follow-up in SIR

In these studies, around two per cent of general intensive care admissions lacked follow-up. The registration of age, comorbidity, location before ICU admission, vasoactive therapy before admission, planned/unplanned admission, surgical status, anatomical site of surgery, reason for admission, length of hospital stay before admission, and acute infection on admission is obligatory so there were no missing data. However, there is no guarantee that entered data are correct, and no check of the quality of SIR data by external controllers has been made to date. In Studies II-III, the percentage of admissions with no physiological parameters missing was 59.1 %, 24.4 % had 1-2 physiological parameters missing and the remaining 16.5% missed 3-10 physiological parameters.

The situation regarding cardiothoracic intensive care data is different, with more planned admissions and less missing data. In Study IV around 1% of admissions were not followed up. Similarly, the number of physiological parameters missing was low for patients admitted to cardiothoracic intensive care after cardiac surgery with ECC, with over 96 % complete data.

The Swedish National Patient Register

During the 1960s, the National Board of Health and Welfare began collecting information regarding in-patients at public hospitals; the National Patient Register (NPR). Since 1987, the NPR includes all in-patient care in Sweden. The quality of data in the NPR has been evaluated and found to be excellent [45,46].

Linking of data

In Study I, the Swedish personal identity number of all patients admitted to intensive care and registered in SIR during the study period was sent to the National Board of Health and Welfare. The Swedish personal identity number of each patient was then linked to the NPR. By matching care intervals in the SIR and the National Patient Register, information on patient hospitalisation and in-hospital mortality became available.

Patients

For the purpose of this thesis, patients were limited to those 16 years or older. A summary of the patients in the studies is shown in Table 4.

Table 4. Patients used in the studies.

Study	I	II	III	IV
Years	2009-2010	2011-2014	2011-2014	2008-2016
Inclusion	General intensive care	General intensive care	General intensive care	Cardiothoracic intensive care
Admissions, total	81,576	143,155	143,163	29,187
Exclusions	Lost to follow-up Readmissions	Lost to follow-up Readmissions	Lost to follow-up Readmissions	Lost to follow-up Missing/no ECC-time
Study cohort	48,861	111,275	107,310	27,814
Age, years (median, IQR)	63 (45 - 75)	65 (47 – 75)	65 (48–75)	67 (60 - 74)
Proportion males (%)	55.9	56.7	56.8	73.4
Length of ICU stay, hours (median, IQR)	22 (12 - 50)	23 (12 – 56)	23 (12–54)	22 (19-53)
ICU mortality %	7.7	8.5	8.3	1.4
30-day mortality %	16.7	17.4	17.5	2.5

ICU = intensive care unit. IQR = interquartile range. ECC-time = extracorporeal circulation time.

Measures of model performance

The performance of a risk adjustment model should be evaluated using several measures [47].

Discrimination

Discrimination in this context describes the model's ability to distinguish between patients who die from those who survive.

Accurate predictions discriminate between those with and those without the specified outcome. Several measures can be used to indicate how well we classify patients in a binary prediction problem. The most common measure of a model's ability to discriminate is the concordance (*c*) index which for a binary outcome is identical to the area under the receiver operating characteristics curve.

Area under the receiver operating characteristics curve (aROC)

The receiver operating characteristics (ROC) curve plots the sensitivity (true positive rate) against $1 - \text{specificity}$ (false positive rate) for all observed threshold settings (Figure 4). The area under the receiver operating characteristics curve (aROC) can have a value between 0.5 and 1, where 1 indicates perfect classification and 0.5 that random classification is equally good. The aROC is a rank-order statistic for correlation between predictions and true outcomes, and as a rank-order statistic, it is insensitive to systematic errors in calibration such as differences in average outcome.

Roughly an aROC ≥ 0.7 is described as good, ≥ 0.8 as very good and ≥ 0.9 is regarded as excellent [48].

Figure 4. Example of area under the receiver operating characteristics curve (aROC).

Calibration

Calibration describes the agreement between observed outcomes and predictions. If the model, for example, predicts a 10% death risk for a particular risk profile, this means that 1 in 10 individuals with this profile die if the model is well calibrated.

Standardised Mortality ratio (SMR)

SMR is simply the number of deceased divided by the predicted number of deceased. A value below 1 indicates that fewer died than expected, which is good when evaluating medical care. When evaluating a model in a dataset used for reference, the SMR does not significantly differ from 1 if the model is well calibrated.

Cox calibration regression

Another measure of calibration is Cox's calibration regression (not to be confused with Cox's semi-parametric method of survival assay, Cox's proportional hazards model). Cox's calibration regression is a regression analysis with expected outcomes as the only input variable and dead / alive as outcome. The intersection of the y-axis (Cox's intercept) and the line coefficient of slope (Cox's slope) can be used to describe the relationship between expected and observed mortality [49].

Hosmer-Lemeshow goodness of fit test

A common way to formally assess the accuracy of calibration is to compare prediction with outcome groupwise using Hosmer-Lemeshow's goodness of fit test (Hosmer-Lemeshow's C and H test respectively). The test is a chi-square test where the patients are grouped according to quantiles of predicted mortality rate (H-test, Figure 5) or quantiles of number of patients when ordered after predicted mortality rate (C-test). The quantiles are usually ten or twenty. For each group the squared difference between observed and predicted number of deaths is divided by the predicted number of deaths. The procedure is repeated for survivors. The result is summed for the groups and compared to a chi-square distribution. The test is very sensitive to sample size.

Figure 5. Example of Hosmer Lemeshow H-test. Observed and predicted number of deaths in spans of predicted mortality rate. For each group the squared difference between observed and predicted number of deaths is divided by the predicted number of deaths. The procedure is repeated for survivors. The result is summed for the groups and compared to a chi-square distribution.

Calibration plot

Calibration plots show observed mortality drawn per risk range against expected mortality (Figure 6). It provides a good picture of calibration across risk classes but little formal information about the statistical sureness of the analysis.

Figure 6. Example of calibration plot. The observed mortality rate is calculated for quantiles of predicted mortality rate.

Calibration belt

The SMR, Hosmer-Lemeshow statistics, and the Cox's calibration test give no information about calibration across risk classes. The use of calibration plots is one way of displaying calibration across risk classes. Calibration belt is a method used to create a confidence band for the calibration plot based on a function that relates expected to observed probabilities across risk classes (Figure 7). The calibration belt allows identification of ranges of risk where there is a significant deviation from ideal calibration [50,51].

Figure 7. Example of calibration belt.

Overall performance measures, accuracy

Brier's score

Brier's score is an overall measure that describes the information value in risk estimation. It was originally developed in the field of meteorology to assess the accuracy of weather forecasts [52]. The difference between observed and predicted mortality is squared for each patient and after summing the results divided by the number of patients. Brier's score can assume values from 0 for a perfect model to 0.25 for a non-informative model with 50% incidence of actual outcome. If the incidence of outcome is lower, e.g. 18%, the maximum score becomes lower (0.148).

Scaled Briers score

Brier's score can be normalised to obtain an adjusted or "scaled" Brier's score that varies between 0% (non-informative model) and 100% (perfect model). The maximum Brier's score for a predicted mortality rate (p) is: $p^*(1-p)$. It may be easier to compare models with different rates of mortality this way [47].

Pseudo R2

R2 is the amount of variation that is explained by a linear model. Explained variation (R_e) is the most common performance measure for continuous outcomes. In logistic regression, it is called pseudo-R2. Its value is difficult to interpret, and its use is debated.

Confidence intervals

Confidence intervals of SMR

When SMRs are used and the number of deaths is low, care should be taken to take the confidence intervals into account and not treat differences caused by random variation as differences in quality of care. Confidence intervals of SMR cannot be calculated directly by standard statistical formulae.

Exact confidence intervals for the SMR is found by first finding the lower (u_L) and upper (u_U) confidence limits of the Poisson distributed observed number of deaths (D) and then calculating the lower $SMR_L = \frac{u_L}{E}$ and upper $SMR_U = \frac{u_U}{E}$ confidence limits of the SMR. For confidence limits of SMR when the number of observed deaths is above 20, Byar's approximation is sufficiently correct [53].

$$u_L = D \left(1 - \frac{1}{9D} - \frac{Z}{3D^{1/2}} \right)^3$$

$$u_U = (D + 1) \left(1 - \frac{1}{9(D+1)} + \frac{Z}{3(D+1)^{1/2}} \right)^3$$

Z denotes the 100(1-alpha/2) percentile of normal distribution

D=observed number of deaths

E=expected number of deaths

Bootstrapping

Bootstrapping is a method that relies on random sampling with replacement. Bootstrapping allows calculation of measures of accuracy such as variance and confidence intervals. It allows calculation of these measures without relying on the assumption of a parametric model, i.e. that values are normally distributed. It is often used when data are not normally distributed and can be used when parametric inference is impossible or requires complicated formulae for calculation of standard errors. It can also be used for hypothesis testing. The method is simple but computationally demanding. To calculate variance a number of samples equally as large as the study cohort is randomly drawn from the study cohort. This means that some patients will show up more than once, while others not at all. This is repeated a large number of times (i.e. 1000) and the measures of interest calculated for each of these resampled cohorts. The variance of the measures calculated for the resampled cohorts is used to calculate, for example, confidence intervals (Figure 8).

Figure 8. Histogram explaining bootstrapping. The simplest way to find the 95 % confidence intervals of SMR in this example with 1000 bootstrap samples is to order the SMRs from lowest to highest and get the value of the number 25 and number 975. There are more exact ways.

Funnel plots

Funnel plots is a method of presenting data which draws attention to the influence that the number of occurring events has on the precision of the measurement (Figure 9). They are essentially a form of control chart where the observed outcome measure is plotted against the number of events and the control limits of when a departure from the expected is so big that it cannot be explained by random variation form a funnel around the target outcome. These plots were first used within meta-analyses to determine to what extent the size of trials influenced the range of their results. Funnel plots have then been used to investigate hospital outcomes, such as risk adjusted postoperative mortality rates [54].

Figure 9. Example of funnel plot. Each point represents a ward and the lines represent 95 % and 99 % confidence intervals.

Missing data

Importance of missing data

If data are missing but loss does not happen at random, then results may be biased. The three main problems with missing data are: 1. that it may introduce bias; 2. it makes analysis of data more complicated; and 3. it reduces efficiency.

Patterns of missing data.

Missing data are commonly classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [55]. When data are MCAR, loss does not depend on observed or unobserved data, patient records with complete data cannot be distinguished from those with incomplete data. Indirect evidence of MCAR can be formally evaluated using Little's test, which if significant rejects the possibility of MCAR [56]. A weaker condition for a missing data mechanism, MAR, is when loss of data depends on values in the observed data but is independent of unobserved data. In this case the pattern of loss may be predicted from non-missing variables in the dataset. MNAR is when loss of data depends also on unobserved data, and cannot be predicted from other variables in the dataset [55].

Imputation

Imputation is the process of replacing missing data with substituted values. Other, known parameters of an admission are used to predict the values of the missing parameter. The known data in the study cohort is used for this prediction.

Generally, cases with any missing data are discarded in multivariable statistics which may introduce bias or affect the representativeness of the results. Imputation preserves all cases by replacing missing data with estimated values based on other available information. When all missing values have been imputed, the dataset can then be analysed using standard techniques for complete data.

Predictive mean matching

Hot-Deck Imputation replaces missing data with comparable data from the same dataset. A widely used method for generating hot-deck imputations is predictive mean matching (PMM), which imputes missing values by means of the nearest-neighbour donor with distance based on the expected values of the missing variables conditional on the observed covariates [57].

Multiple imputation

To account for increased noise due to imputation, the imputation process may be repeated to give different imputed values of the missing data and return a number of slightly different complete datasets. After analysis of the complete datasets, the results are pooled into one result by calculating the mean, variance, and confidence interval of the variable of concern using the method Rubin developed [58].

Multiple imputation can be used in cases where data are missing completely, missing at random, and even when the data are missing not at random.

Model validation

Evaluation of the predictive performance of a model requires other data than the model was developed with. A model may work very well using the dataset from which it was developed but perform badly with a new dataset. Performance should therefore be evaluated using other datasets.

Recalibration of models

A statistical model may be recalibrated in different ways. If miscalibrations are similar in different subsets of patients, simple recalibration of overall coefficients may be enough. Usually the two coefficients for slope and intercept and in some cases correction of the effect of outliers are used. Simple recalibration does not risk fitting coefficients too well in the current dataset to be able to perform well in other datasets, but may not improve calibration in different quantiles of risk as much. Simple recalibration does not improve the relation of calibration in different groups of patients (the uniformity of fit) that much.

If there are substantial differences in subsets of patients, a full recalibration of all coefficients may be necessary.

Model development

Regression

Most current prognostic models in medicine are developed using logistic regression. There are many other ways than plain linear or logistic regression for model building. For example, neural networks and support vector machines. They can generate good models, but they are not easily explained and thus transparency decreases. While they generally perform somewhat better, they are more complex to understand and to implement. To explain regression, it is easier to start with linear regression.

Linear regression

Linear regression (Figure 10) is a linear approach for modelling the relationship between a scalar-dependent outcome variable and one or more explanatory variables (or independent variables). When there is more than one explanatory variable, the process is called multivariable linear regression. In multivariate linear regression, multiple correlated dependent variables are predicted rather than a single scalar variable.

Figure 10. Example of simple linear regression, which has one independent variable.

Logistic regression

Logistic regression uses one or more predictor variables that may be either continuous or categorical. Unlike linear regression, logistic regression is used for predicting outcomes that take membership in one of a limited number of categories rather than a continuous outcome. For this, a way to convert a binary variable into a continuous one is needed. To do that, logistic regression first takes the odds of the event happening for different levels of each predictor variable. Next, takes the ratio of those odds (which is continuous but cannot be negative) and then takes the logarithm of that ratio to create a continuous transformed version of the dependent variable (Figure 11). Logistic regression was developed by statistician David Cox in 1958 [49].

Multivariate logistic regression predicts more than one outcome variable while multivariable logistic regression uses more than one predictor variable.

Figure 11. The result of the logistic function that is used to transform the ratio of the odds of the event to a continuous variable may vary between minus infinity and infinity.

Cubic splines

Cubic splines can be used to transform a non-linear predictor into a linear increase in risk. A spline is in mathematics a special function and is in this case used to describe the shape of a curve. Cubic spline regression fits cubic functions that are joined at a series of knots. A plot of these functions will look smooth if they have the same first and second derivatives at the knots. The fact that the spline functions are linear in the regression coefficients means that we can use standard methods of inference [59]. Cubic splines tend to behave poorly at the two tails (before the first knot and after the last knot). To avoid this, restricted cubic splines may be used (Figure 12). A restricted cubic spline is a cubic spline in which the splines are constrained to be linear in the two tails. This generally provides a better fit to the data and has the effect of reducing the degrees of freedom.

Figure 12. Example of restricted cubic spline. The points show the mortality rate for different age of patients. The blue line is the mortality risk described with a restricted cubic spline with five knots.

Overfitting

If a model is developed from a database that is small with too few occurrences of the outcome of interest, the individual coefficients of the model are likely to become too large. The model performs well with the development dataset, but performance deteriorates when used with a different dataset. This is called overfitting and leads to poor performance when the model is used with other datasets.

Penalisation

When there is risk of overfitting, penalisation can be used to shrink the individual coefficients of the model towards zero. Several forms of penalisation exist, each having its own benefits. Ridge regression can only shrink coefficients towards zero, Lasso regression can also perform variable selection, and elastic net regression performs better if variables are highly correlated [60–62].

Methods specific to each study

Study I

Customisations were made for the SAPS3 model using 30-day and in-hospital mortality as outcome. Any differences in the survival distribution between observed mortality rate at hospital discharge (OMR_{IH}) vs. observed mortality rate 30 days after ICU admission (OMR_{30d}) were analysed using McNemar's method[63]. The unconditional profile likelihood method [64] was used to calculate 95% CI for differences between OMR_{IH} and OMR_{30d} . Byar's method [53] was used to determine 95% CI for 30-day SMR (SMR_{30d}) and in-hospital SMR (SMR_{IH}). We used I2 statistics on SMR_{IH} and SMR_{30d} within each case-mix category to describe the percentage of variation within each category that was related to group heterogeneity, rather than chance [65,66].

P-values for differences between SMR_{IH} and SMR_{30d} for each ICU and for each case-mix group were calculated using bootstrapping with 10,000 replications. A p-value < 0.05 was generally chosen as level for statistical significance, but where multiple comparisons were performed, the significance level $p < 0.001$ was also presented.

Study II

We used admissions during 2011-2012 to develop the 30-day, 90-day and 180-day models ($SAPS_{30d}$, $SAPS_{90d}$ and $SAPS_{180d}$). Level 1 recalibra-

tions were made for all models i.e. recalibrating the overall coefficients (rather than the coefficients that make up the SAPS3 score) transforming the score to a predicted mortality rate. In this way the predicted overall mortality as well as the predicted mortality in different risk quantiles equals the observed mortality. To do this we did the following: 1) we used the original SAPS3 admission score function [$\text{Logit} = x + \ln(\text{SAPS3 score} + y) * z$, probability of death = $e^{\text{Logit}} / (1 + e^{\text{Logit}})$]; 2) we optimised y to minimise the skew of $\ln(\text{SAPS3 score} + y)$ and attain a normal distribution, and 3) we recalculated x and z to recalibrate the model. This procedure was carried out on the development cohort using 30-day, 90-day and 180-day mortality, respectively, as outcome measures.

Byar's method, as outlined by Breslow and Day [53], was used to determine SD, SE of the mean and CIs for SMR.

Study III

To understand the magnitude of the effect of loss of physiological data, we simulated missing data. We used the full study cohort and imputed missing physiological parameters (six imputed study cohorts) using predictive mean matching with the mice package in R [67]. For each physiological parameter and for oxygenation and pH together, all values in the imputed datasets were removed and the score and corresponding mortality risk was recalculated. The presented results are mean values for the six imputed study cohorts.

Indirect evidence of MCAR was formally evaluated using Little's test [56], which if significant rejects the possibility of MCAR.

The Mann-Whitney U test was used to compare the numbers of missing data between survivors and non-survivors. CIs for scaled Brier's score, aROC, and SMR were calculated using bootstrapping with 1,000 replications, as were p-values.

Study IV

The development of a model for cardiothoracic intensive care began with variables similar to the ones used by Higgins [35]. Univariate coupling with 30-day mortality was tested for the candidate variables using admissions during 2008 – 2014. Variables with univariate coupling ($p < 0.05$) were retained for model development. We evaluated four different approaches of transforming retained variables into a risk score resulting in four different models. In the first model (modified Higgins' model) we used Higgins' original cut-offs except for oxygen extraction fraction (O_2EF) and base excess (BE) which were not included in the original model [35]. In the second model (Youden model) we identified new cut-

offs using the Youden index [68]. In the third model we used restricted cubic splines with four knots to transform variables into rates of risk increase (restricted cubic spline model). In the fourth model multiple cut-offs were analysed generating groups which were deemed reasonable from clinical experience, loess smoothed plots [69] and number of admissions per group (multiple cut-offs model). Groups were merged in the multiple cut-offs model if the effect on risk was not significant ($p < 0.05$) or if the effects on risk for adjacent groups were similar. Missing variables were imputed using predictive mean-matching in the development of the restricted cubic spline model and by assigning values associated with minimal risk (zero points in the multivariable model) in the modified Higgins model, the Youden model and the multiple cut-off model.

Variables were kept in the model development if the p -value was < 0.05 . The variation inflation index was used to check for collinearity between variables [70]. Significant interactions were searched for between central venous pressure (CVP), heart rate (HR) and O_2EF since interactions could be anticipated. Penalisation using elastic net regression [62] was used to shrink the individual coefficients towards zero so as to reduce the risk of overfitting the model to the development dataset.

If calibration was non-linear, a shrinking power transformation was used [24]. This procedure uses log-transformation of the score to reduce the influence of outliers on the mortality prediction. For this purpose, the transformed $\log(\text{score} + g)$ was used to predict 30-day mortality, where g was chosen to minimise the skewness of $\log(\text{score} + g)$.

The final multivariable model was evaluated on admissions during 2015-2016 which were not accessed during the model development process.

Changes in parameters over time were evaluated with linear or logistic regression.

Software packages

R (version 3.0.3 - 3.4.2; R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria) and STATA/SE (version 13.0 - 15.0; StataCorp LP, College Station, TX) were used for all analysis and graphical presentation.

RESULTS

The performance of SAPS3 in Swedish intensive care.

When evaluated using data from 35,610 admissions 2009 – 2010, the discrimination of SAPS3 was good having an aROC of 0.86 when predicting both in-hospital and 30-day mortality. Calibration using the original (general) SAPS3 formula for predicting in-hospital mortality was poor with an SMR of 0.63, Cox's intercept -1.06 and Cox's slope 1.11. Equally poor calibration was found when using the original (general) formula for predicting 30-day mortality with an SMR of 0.65, Cox's intercept -0.79 and Cox's slope 1.08 (Study I, partly unpublished data).

Good discrimination implies that SAPS3 may well be useful in Sweden, but that recalibration is needed to achieve satisfactory performance. This is necessary not only to bring SMRs for Swedish ICUs closer to 1, but also to lessen the effect of lack of calibration in the various risk-classes on SMR. A simple recalibration was chosen for stability while retaining the possibility to present the SAPS3 score and estimated mortality using the original (general) SAPS3 formula.

Recalibration of the model improved its calibration greatly as shown in Figure 13.

Figure 13. Calibration of original (general) and recalibrated SAPS3 model when predicting 30-day mortality for 57,729 admissions during 2013-2014.

The effect of missing data on performance of the model.

Although a considerable amount of physiological data was missing, data describing patient characteristics prior to admission and the circumstances of admission were complete. 59.1% of records were complete (Table 5). Overall, 11% of physiological data were missing.

There were important differences in the amount of physiological data missing; oxygenation was most frequently missed followed by bilirubin and pH (Table 5). Missing data were more frequent in survivors than in non-survivors (on average, 1.15 vs. 0.95 missing data per patient, $p < 0.001$). The average number of missing data decreased with increasing risk: lowest-risk quartile, 1.65; second-lowest-risk quartile, 1.31; third-lowest-risk quartile, 0.92; and highest-risk quartile, 0.55 missing data per patient.

Table 5. Percentage of records with physiological variables missing.

Variables	Number of admissions	Proportion of study cohort, %
0 missing	63,392	59.1
1 missing	15,279	14.2
2 missing	10,982	10.2
3 missing	5,419	5.0
4 missing	4,671	4.4
5 missing	2,111	2
6 missing	3,479	3.2
7 missing	948	0.9
8 missing	293	0.3
9 missing	546	0.5
10 missing	190	0.2
Missing variable ^a		
Bilirubin	22,922	21.4
Oxygenation	26,868	25
Temperature	6,765	6.3
pH	19,831	18.5
GCS or RLS85	2,335	2.2
Platelets	12,786	11.9
Creatinine	11,202	10.4
Leukocytes	12,317	11.5
Heart rate	2,081	1.9
Blood pressure	2,300	2.1

GCS Glasgow Coma Scale, RLS85 Reaction Level Scale 85. ^a The sum of missing variables is greater than 100% because more than one variable may be missing.

The impact of missing data on performance was examined for increasing numbers of missing data and per variable in the original dataset, as shown in Table 6. Mortality differed for subsets of different missing variables, indicating that data were not missing completely at random. This notion was supported by Little's test, which was significant ($p < 0.001$). SMR showed poor calibration for datasets with more than one missing value.

Table 6. Model performance with increasing number of missing variables and with specific variables missing.

Missing variables	Patients	aROC	SMR
Study cohort	107,310	0.85 (0.85 - 0.86)	0.99 (0.98 - 1.01)
Number of missing variables			
0 missing	63,392	0.84 (0.84 - 0.85)	0.92 (0.90 - 0.93)
1 missing	15,279	0.87 (0.86 - 0.87)	1.05 (1.02 - 1.08)
2 missing	10,982	0.89 (0.88 - 0.90)	1.10 (1.05 - 1.15)
3 missing	5,419	0.89 (0.87 - 0.90)	1.18 (1.11 - 1.25)
4 missing	4,671	0.87 (0.85 - 0.88)	1.27 (1.20 - 1.34)
5 missing	2,111	0.87 (0.86 - 0.89)	1.50 (1.38 - 1.61)
6 missing	3,479	0.87 (0.86 - 0.89)	1.70 (1.58 - 1.82)
7 missing	948	0.91 (0.89 - 0.94)	2.03 (1.83 - 2.24)
8 missing	293	0.85 (0.79 - 0.91)	2.69 (2.13 - 3.24)
9 missing	546	0.88 (0.85 - 0.92)	2.71 (2.30 - 3.13)
10 missing	190	0.76 (0.70 - 0.84)	4.94 (3.73 - 6.14)
Specific missing variables			
Bilirubin	5,163	0.85 (0.84 - 0.86)	1.03 (0.98 - 1.08)
Oxygenation	6,066	0.88 (0.87 - 0.90)	1.08 (1.02 - 1.14)
Temperature	1,519	0.85 (0.83 - 0.87)	1.11 (1.02 - 1.19)
pH	656	0.90 (0.87 - 0.95)	0.80 (0.62 - 0.98)
GCS or RLS85	1,015	0.78 (0.75 - 0.82)	1.08 (0.97 - 1.21)
Platelets	338	0.84 (0.79 - 0.88)	1.08 (0.91 - 1.25)
Creatinine	159	0.91 (0.86 - 0.95)	1.04 (0.80 - 1.27)
Leukocytes	249	0.88 (0.83 - 0.93)	0.88 (0.69 - 1.06)
Oxygenation and pH	7,519	0.89 (0.88 - 0.91)	0.97 (0.90 - 1.03)

When specific variables are shown, only these variables are missing and all others are present. Missing values are presumed normal and given 0 points, as in the original SAPS3 model. Performance data when more than six variables are missing are only shown for completeness, and data for missing heart rate ($n = 31$) or blood pressure ($n = 83$) are not shown. aROC = area under the receiver operating characteristics curve, GCS = Glasgow Coma Scale, RLS85 = Reaction Level Scale 85, SAPS3 = Simplified Acute Physiology 3, SMR = standardised mortality ratio.

Table 7 shows the impact on model performance of simulated compared with originally missing data. Simulation of one missing variable at a time had only a minor impact on the aROC for most variables. For oxygenation, pH and a combination of both, performance was worse with simulated missing data than with the originally missing data. For cerebral responsiveness the effect was the opposite. The effect on SMR was also minor except for oxygenation, pH and a combination of both where it increased when they were simulated as missing. A full dataset where the missing values were imputed was used before simulating missing data. The rationale for this was to have an unbiased dataset to compare the datasets with originally missing data.

Table 7. Model performance in datasets with primary and simulated missing variables. The full study cohort is shown for comparison in the first row. The second row shows the full study cohort with missing variables imputed.

Missing variable(s)	Dataset	aROC (95%CI)	SMR (95%CI)	Changed score, %
Study cohort	Primary	0.85 (0.85 - 0.86)	0.99 (0.98 - 1.01)	
Study cohort	Imputed	0.86 (0.85 - 0.86)	0.96 (0.95 - 0.97)	
Bilirubin	Primary	0.85 (0.84 - 0.86)	1.03 (0.99 - 1.08)	
	Simulated	0.86 (0.85 - 0.86)	0.98 (0.97 - 0.99) *	16.8
Oxygenation	Primary	0.88 (0.87 - 0.90)	1.08 (1.02 - 1.14)	
	Simulated	0.85 (0.85 - 0.85) ***	1.18 (1.17 - 1.20) ***	41.3
Temperature	Primary	0.85 (0.83 - 0.87)	1.11 (1.02 - 1.19)	
	Simulated	0.86 (0.85 - 0.86)	0.99 (0.98 - 1.00) **	16.2
pH	Primary	0.90 (0.87 - 0.95)	0.80 (0.62 - 0.98)	
	Simulated	0.86 (0.85 - 0.86) *	1.00 (0.98 - 1.01) *	24.6
GCS or RLS85	Primary	0.78 (0.75 - 0.82)	1.08 (0.97 - 1.21)	
	Simulated	0.84 (0.83 - 0.84) **	1.13 (1.11 - 1.14)	37.1
Platelets	Primary	0.84 (0.79 - 0.88)	1.08 (0.91 - 1.25)	
	Simulated	0.86 (0.85 - 0.86)	0.99 (0.98 - 1.00)	17.9
Creatinine	Primary	0.91 (0.86 - 0.95)	1.04 (0.80 - 1.27)	
	Simulated	0.85 (0.85 - 0.86) *	1.07 (1.05 - 1.08)	36.4
Leukocytes	Primary	0.88 (0.83 - 0.93)	0.88 (0.69 - 1.06)	
	Simulated	0.86 (0.85 - 0.86)	0.99 (0.98 - 1.00)	32.4
Oxy and pH	Primary	0.89 (0.88 - 0.91)	0.97 (0.90 - 1.03)	
	Simulated	0.85 (0.84 - 0.85) ***	1.24 (1.23 - 1.25) ***	45.9

The number of records in the primary datasets with missing variables are shown in Table 3, while the number in simulated datasets are the same as in the full study cohort ($n = 107,310$). Simulated datasets show mean values from six imputed study cohorts (full cohorts with missing values imputed, see Methods) where all values of the variable(s) selected were removed. Changed score denotes percentage of records where the variable was given 0 points; the balance had already 0 points (variable was within normal limits). *, ** and *** denote $p < 0.05$, < 0.01 , and < 0.001 , respectively, compared with primary datasets where the same variable(s) was missing. Oxy and pH denotes that both oxygenation and pH were missing. aROC area under the receiver operating characteristics curve, GCS = Glasgow Coma Scale, RLS85 = Reaction Level Scale 85, SAPS3 = Simplified Acute Physiology Score 3, SMR = standardised mortality ratio.

The relationship between 30-day and in-hospital mortality.

In the study cohort the overall 30-day mortality was 2.6 percentage points higher (95% CI 2.2-3.1%) than in-hospital mortality (16.7% respective 14.1%). The difference increased with age and severity of illness and was notably largest in those discharged from ICU to another hospital. This is described in Study I.

The use of a time fixed mortality endpoint instead of in-hospital mortality for risk adjustment.

As described in study I, SAPS3 recalibrations were made for in-hospital (IH) and 30-day (30d) mortality. SMR_{30d} (95%CI) and SMR_{IH} (95%CI) were 1.00 (0.97-1.03) and 1.00 (0.97-1.03) respectively.

There was heterogeneity ($I^2 > 60\%$) across the 54 ICUs for both SMR_{IH} and SMR_{30d} , and for ten ICUs SMR_{30d} was significantly higher than SMR_{IH} . For both SMR_{IH} and SMR_{30d} there was considerable heterogeneity ($I^2 > 80\%$) across case-mix groups for age, severity of illness, surgical status, diagnosis, length of hospital stay and discharge destination (Figure 14). SMR_{30d} and SMR_{IH} differed significantly for elective surgery, sepsis, in all hospital length of stay (LOS) subgroups and in subgroups with different discharge destinations (Figure 14).

Figure 14. In-hospital standardised mortality ratio (SMR_{IH}) and 30-day standardised mortality ratio (SMR_{30d}) $\pm$ 95% CI for case-mix groups defined by age, SAPS3 score, surgical status at admission, primary ICU ICD-10 diagnosis, length of hospital stay and discharge destination. * SMR_{30d} differs from SMR_{IH} ($p < 0.05$). *** SMR_{30d} differs from SMR_{IH} ($p < 0.001$). Other ICU can be at the same hospital. Observe that the in-hospital mortality is zero for patients discharged to a ward at another hospital.

The calibration plots for in-hospital and 30-day mortality are similar (Figure 15), which also indicates that the SAPS3 model performs well with 30-day mortality as outcome.

Figure 15. Calibration plots for the customised SAPS3 in-hospital model and SAPS3 30-day model.

The usefulness of the SAPS3 outcome prediction model using different time-fixed outcomes.

Results of level 1 customisation for all three models are shown in Table 8. Discrimination for models predicting 30-day, 90-day and 180-day mortality as outcomes on the validation dataset was good (aROC = 0.85, 0.84, and 0.83, respectively). The uniformity of fit for all patients as shown by SMR was good (not significantly different from 1.00) for all three models. The scaled Brier's score was equivalent across the three fixed times, as were the results from Cox's calibration regression. The latter indicated good fit, reinforcing the SMR results.

Table 8. Performance measures of the models in the validation dataset.

Model	SAPS3 _{30d}	SAPS3 _{90d}	SAPS3 _{180d}
aROC	0.85	0.84	0.83
SMR (s.e.)	0.99 (0.01)	0.99 (0.01)	1.00 (0.01)
Brier's score	0.11	0.12	0.13
Scaled Brier's score	0.27	0.28	0.27
Cox's intercept	- 0.06	- 0.04	- 0.04
Cox's slope	0.96	0.96	0.95
H-L C-test χ^2			
Df=8	61.6	68.4	77.5
H-L H-test χ^2			
Df=8	66.4	63.7	81.4

Values are means. SMR = standardised mortality ratio, s.e. = standard error of the mean, H-L = Hosmer-Lemeshow, Df= degrees of freedom, aROC = area under the receiver operating characteristics curve.

Figure 16 shows the calibration plot for all three models, respectively. Each model showed close agreement between mean observed and mean predicted mortality across deciles of predicted mortality.

Figure 16. Calibration plots for the models in the validation dataset (2013-2014). The points overlap each other, the 30-day model can hardly be seen in the plot.

Effect of the outcome measure used when comparing intensive care units.

Funnel-plots of SMR for individual ICUs showed that 7 ICUs (11.1%) moved across the 95 % CI boundaries when follow-up was extended from 30 to 90 days and 8 units (12.7%) when extended from 30 to 180 days after ICU admission (Figure 17).

Figure 17. Funnel plots of SMR (from top to bottom: SMR_{30d}, SMR_{90d} and SMR_{180d}) calculated from the external validation dataset collected 2013-2014. Individual ICUs are identified by numbers. Dashed lines 95% confidence intervals.

Tables 9a and 9b summarise the comparison of SMRs between the 30-day model and the 90- or 180-day models. The SMRs for the 30-day model were significantly ($p < 0.05$) less than 1.00 in 12/63 (19%) ICUs, significantly ($p < 0.05$) greater than 1.00 in 11/63 (17%) ICUs, and close to 1.00 in 40/63 (63%) ICUs. No ICU went from significantly less/greater than 1.00 to significantly greater/less than 1.00 in the 90- and 180-day models.

Table 9a. Comparisons of SMRs for individual ICUs based on the 30-day (SAPS3_{30d}) and 90-day models (SAPS3_{90d}). ICUs were grouped depending on whether the limits of 95 % confidence interval of SMR were below, above or included 1.

	SAPS3_{90d} SMR < 1.00	SAPS3_{90d} SMR ~ 1.00	SAPS3_{90d} SMR > 1.00
SAPS3_{30d} SMR < 1.00	9	3	0
SAPS3_{30d} SMR ~ 1.00	1	38	1
SAPS3_{30d} SMR > 1.00	0	2	9

Table 9b. Comparison of SMRs for individual ICUs based on the 30-day (SAPS3_{30d}) and 180-day models (SAPS3_{180d}). ICUs were grouped depending on whether the limits of 95 % confidence interval of SMR were below, above or included 1.

	SAPS3_{180d} SMR < 1.00	SAPS3_{180d} SMR ~ 1.00	SAPS3_{180d} SMR > 1.00
SAPS3_{30d} SMR < 1.00	8	4	0
SAPS3_{30d} SMR ~ 1.00	1	38	1
SAPS3_{30d} SMR > 1.00	0	2	9

Risk prediction after cardiovascular surgery.

We found that the performance of the in Sweden currently used customisation of Higgins-ICU had deteriorated and opted for development of a new model. To develop the model, we used the variables which are collected by cardiothoracic ICUs affiliated to SIR. Details of the candidate variables including proportion of values missing are shown in Table 10. Data were complete for more than 96 % of admissions.

Three candidate variables (number of previous heart operations, previous vascular surgery and gender) were not associated with outcome in the univariate analysis and thus excluded. We did not find any significant interactions between CVP, HR and O₂EF.

Compared to the modified Higgins' model, discrimination improved by optimising cut-offs using the Youden index and by using multiple cut-offs or restricted cubic splines. The overall calibration assessed by Cox's calibration regression was perfect as expected for all models using the development dataset. However, all models performed worse when examined using the Hosmer-Lemeshow C test and calibration belts (Table 11). Log-transformation resulted in better calibration in quantiles of predicted mortality rate for all models except for the modified Higgins' model. The log-transformed multiple cut-off model performed like the log-transformed cubic splines model, the latter having missing values imputed. The effect of shrinkage on performance in the development dataset and on coefficients was small. After shrinkage the coefficients were rounded to two decimals and performance was found to be similar.

Table 10. Candidate variables.

Variable	Median (iqr)	Missing
Albumin, g/l	38 (35 - 40)	0.48 %
Age, years	67 (59 - 74)	0.00 %
Number of previous heart operations	0 (0 - 0)	0.03 %
Arterial oxygen saturation, %	98 (97 - 99)	0.07 %
Arterial carbon dioxide partial pressure, kPa	5.3 (4.9 - 5.8)	0.04 %
Arterial oxygen partial pressure, kPa	14.4 (11.8 - 18.3)	0.04 %
Base excess, mmol/l	-1 (-3 - 0)	0.08 %
Mixed venous saturation ^a , %	65 (60 - 70)	60.9 %
Central venous saturation ^a , %	68 (63 - 73)	40.8 %
CVP, mmHg	8 (5 - 10)	1.20 %
ECC-time, min	92 (69 - 128)	0.00 %
Heart rate, beats/min	78 (70 - 83)	0.04 %
Inspired oxygen fraction	45 (40 - 50)	0.03 %
Creatinine, umol/l	84 (73 - 99)	0.08 %
Height, m	174 (167 - 180)	0.14 %
Weight, kg	80 (71 - 90)	0.08 %
BSA, m ²	2.0 (1.8 - 2.1)	0.14 %
Oxygen extraction, mixed venous ^a , %	33 (29 - 38)	61.0 %
Oxygen extraction, cantral venous ^a , %	30 (26 - 36)	40.9 %
A-a gradient, kPa	20.4 (15.9 - 28.0)	0.47 %
IABP	0.02	0.01 %

CVP = central venous pressure. ECC-time = extracorporeal circulation time. FiO₂=Inspired oxygen fraction. BSA = body surface area. Aa-gradient = alveolar - arterial oxygen gradient. IABP=Intra-aortic balloon pump or ventricular assist.
^aEither mixed venous saturation or central venous saturation was used. Mixed venous saturation was preferred in cases where both were recorded.

Table 11. Performance of the different models.

Model	aROC (95% CI)	P-value	Intercept	Slope	HL C	Calibration belt
<i>Development cohort</i>						
Modified Higgins	0.83 (0.81-0.85)		0	1	0.031	<0.001
Modified Higgins log	0.83 (0.81-0.85)		0	1	0.003	<0.001
Youden	0.85 (0.83-0.86)	0.006 ^a	0	1	0.023	<0.001
Youden log	0.85 (0.83-0.86)		0	1	0.035	0.093
Restricted cubic splines	0.86 (0.85-0.88)	<0.001 ^a 0.001 ^b	0	1	0.066	<0.001
Restricted cubic splines log	0.86 (0.85-0.88)		0	1	0.418	0.799
Multiple cut-off	0.86 (0.85-0.88)	<0.001 ^a 0.65 ^c	0	1	0.005	<0.001
Multiple cut-off log	0.86 (0.85-0.88)		0	1	0.753	0.695
<i>Validation cohort</i>						
Multiple cut-off log	0.86 (0.83-0.89)		-0.16 (-0.47-0.19)	1.01 (0.89-1.13)	0.293	0.114

aROC = area under the receiver operating characteristics curve. CI = confidence interval. CI was calculated using bootstrapping. Intercept and slope from Cox's calibration regression, HL C denotes Hosmer-Lemeshow C-statistic. For definition of models see Material and Methods. Log denotes that the score was log-transformed as described in Materials and Methods. P-value denotes differences in aROC between models: ^a compared to the modified Higgins model, ^b compared to the Youden model and ^c compared to the restricted cubic splines model.

Although the log-transformed cubic splines and multiple cut-off models performed almost identically, we selected the less complex log-transformed multiple cut-off model as the final model and examined its performance in the validation dataset. Discrimination of the final model was good in the both the development- and the validation dataset (aROC 0.86 in both). Calibration (figure 18) in the validation dataset showed non-significant HL-C statistic and calibration belt, Cox's regression intercept -0.16 (95% CI: -0.47 to 0.19) and slope 1.01 (95% CI: 0.89 to 1.13), while SMR was 0.86 (0.83 – 0.89). The calibration belts for the development and validation datasets are shown in Figure 18.

Figure 18. Calibration belt for the final model. Development dataset in the left panel and validation dataset in the right panel.

DISCUSSION

The performance of SAPS3 in Swedish intensive care.

The good discrimination of SAPS3 in Sweden, even though used with 30-day mortality as outcome when developed for in-hospital mortality, indicates the robustness of SAPS3. An aROC of 0.86 in this study is similar to when SAPS3 was first developed (0.848)[24]. It also indicates good data recording and assessment of reason for admission.

The low SMR in Sweden could be interpreted as Swedish intensive care being much better than European care was when SAPS3 was developed. Calibration of the model, however, is affected by many other factors such as which patient categories are admitted to the ICU. It should be remembered that the SMR was low despite the fact that SAPS3 used 30-day mortality in Sweden whereas it was developed for in-hospital mortality which in our patients was lower than 30-day mortality.

When SAPS3 was recalibrated and evaluated using data from different years, the results were found to be similar, suggesting a stable model. The original SAPS3 formula and the northern European customisation which was also published in the original SAPS3 paper [24] has poor calibration in Sweden (Figure 13). Whether this was caused by a different case-mix than the original SAPS3 cohort or the result of improved care is not known, and we had no idea whether calibration of the recalibrated formula would deteriorate rapidly. It was therefore reassuring to see that calibration remained good in the validation dataset from admissions that were more recent than the dataset we used for recalibration (Figure 16,21). The recalibration of the model improved calibration as shown in the calibration plots (Figure 19).

Figure 19. Calibration plots for original (general) and recalibrated models used for prediction of 30-day mortality.

This can also be shown as the SMR being more even for different spans of predicted mortality rate (Figure 20). This decreases the effect of case-mix on SMR.

Figure 20. Standardised mortality ratio (SMR) for different spans of predicted 30-day mortality before and after recalibration.

The recalibrated SAPS3 model has remained stable over time in Sweden (Figure 21).

Figure 21. The recalibrated Standardised mortality ratio (SMR) based on 30-day mortality has remained stable over time. One year running means are shown.

The uniformity of fit in different subgroups is good but not perfect (Table 12). The differences in SMR in different subgroups could be improved by total recalibration of the SAPS3 model, i.e. calculating new coefficients for all variables included. Total recalibration, however, makes it impossible to compare Swedish SAPS3 scores with international ones.

Table 12. Uniformity of fit of recalibrated models for different subgroups (unpublished data).

Model		30-day	90-day
	Number	SMR (95% CI)	SMR (95% CI)
Total	57107	0.994 (0.976 - 1.014)	1.028 (1.011 - 1.046)
Age>80	8335	1.155 (1.114 - 1.197)	1.174 (1.136 - 1.213)
Age<80	47680	0.941 (0.920 - 0.963)	0.981 (0.961 - 1.001)
Male	32253	0.972 (0.947 - 0.997)	1.006 (0.984 - 1.030)
Female	24854	1.026 (0.997 - 1.057)	1.059 (1.031 - 1.087)
Basic and observational	8434	0.835 (0.753 - 0.923)	1.009 (0.931 - 1.091)
Cardiovascular	17139	1.014 (0.987 - 1.041)	1.013 (0.989 - 1.038)
Gastrointestinal	7478	0.851 (0.812 - 0.892)	0.917 (0.879 - 0.957)
Neurological	18754	1.004 (0.974 - 1.034)	1.003 (0.976 - 1.031)
Renal	7428	0.941 (0.904 - 0.979)	0.978 (0.943 - 1.014)
Respiratory	18568	1.035 (1.007 - 1.063)	1.062 (1.036 - 1.089)
Metabolic	10627	0.930 (0.894 - 0.967)	0.966 (0.932 - 1.001)
Trauma	5218	1.019 (0.942 - 1.100)	1.012 (0.943 - 1.085)
Emergency	51519	0.990 (0.971 - 1.010)	1.022 (1.004 - 1.040)
Surgery	12452	0.822 (0.781 - 0.863)	0.903 (0.865 - 0.943)
No surgery	44655	1.032 (1.011 - 1.054)	1.056 (1.037 - 1.076)

SMR = Standardised mortality ratio.

The effect of missing data on performance of the model.

SAPS3 performs well in daily practice despite some missing physiological data, probably because loss of data is not a completely random event.

While missing data is permitted in the SAPS3 model, the amount and effect of lost data is not well described. When changing from a model which included physiological variables and blood chemistry from the first 24 hours on ICU to a model that only included data from one hour prior to and one hour after arrival at the ICU, it was feared that the amount of missing data would increase and the performance of the model deteriorate. Because the SAPS3 model assumes that missing values are normal, increasing loss of physiological variables will deflate mortality risk leading to falsely high SMRs. As expected, the standardised mortality ratio increased with increasing missing physiological data. But the increase was less than expected since there was a correlation between low risk and greater loss of data. The practical implications of this are that we can continue to use a narrow time frame for sampling physiological data. The clinical decision to omit sampling when there is little need for the safety of the patient is not likely to have a negative effect on SAPS3 performance. The amount of missing data is very similar to what was shown for APACHE III despite the shorter time for including physiological data [71]. They also found that the number of missing physiological data was greater among (hospital) survivors than non-survivors.

The use of a time-fixed mortality endpoint versus in-hospital mortality.

Given the problems of discharge bias associated with in-hospital mortality, time-fixed outcomes provide a more objective measure of ICU performance. These problems arise in that in-hospital mortality is severely biased by discharge policies, financial incentives and the number of transfers between hospitals. It may also be biased in communities where long-term acute care or hospice care facilities exist and patients are transferred to these for prolonged care and rehabilitation or palliative care in cases of terminal illness.

Choice of mortality end-point influences SMR. The extent of influence depends on hospital-, ICU- and patient cohort characteristics as well as interhospital transfer rates. All these factors influence the difference between SMR based on 30-day mortality and SMR based on in-hospital mortality.

There appears to be consensus that time-fixed (for example 30-day) mortality is a less biased measure than in-hospital mortality, and time-fixed mortality endpoints are being demanded more and more [27–29]. Time-fixed mortality is not affected by discharge to another hospital or short length of hospital stay. The most commonly used general intensive care risk adjustment models, however, use in-hospital mortality as outcome measure since time-fixed mortality is harder to achieve in most countries [21,24,32]. We have shown that the SAPS3 model works as well with 30-day mortality as it does with in-hospital mortality for Swedish intensive care risk adjustment. For satisfactory calibration, SAPS3 must be recalibrated for use in Sweden, not only for use with 30-day mortality as outcome, but also for use with in-hospital mortality as originally designed.

Effect of the outcome measure used when comparing intensive care units.

Less than 1% of the study cohort remained on the ICU as long as 30 days. This supports the notion that 30-day mortality is an endpoint that is more likely to represent quality of the intensive care and thus a reasonable alternative to in-hospital mortality.

The time-point that is preferred for evaluating mortality after intensive care has not been reported. We have shown that SAPS3 works well with 30-, 90- and 180-day mortality as endpoint. One would expect that 30-day mortality is more influenced by the intensive care whereas 90- and 180-day mortality are more likely to be influenced by chronic illness, frailty, age and care provided on other wards. From the patient perspective a long-term outcome is of more interest than a short-term, but if the primary goal is to evaluate the quality of intensive care, 30-day mortality is probably more specific. When we compared which ICUs that had SMRs significantly different from 1.00 calculated with 30-, 90-, and 180-day mortality as outcome respectively, they generally agreed. This is no surprise since those patients who die usually do so within 30 days. However, a few units deviated.

Risk prediction after cardiovascular surgery.

Our model was a modification of the Higgins ICU admission model (Table 13, 14). The model requires only 11 recorded variables easily available on admission to the ICU.

Table 13. Original Higgins ICU admission model and models used in Sweden.

	Higgins 1997	Gothenburg 2002	SIR 2010
Body surface area	< 1,72 m ²	< 1,72 m ²	< 1,71 m ²
Prior heart operation	1 2 or more	1 2 or more	
Prior vascular intervention	Y/N	Y/N	
Age	>= 70 years	>= 70 years	> 71 years
Preoperative creatinine	> 1,9 mg/dl (=168 umol/l)	> 168 umol/l	> 165 umol/l
Preoperative albumin	< 3,5 mg/dl (=35g/l)	< 35 g/l	< 35 g/l
CPB time	>= 160 min	>= 160 min	> 153 min
Use of IABP after CPB	Y/N	Y/N	IABP/VAD Y/N
A-a O ₂ gradient	>=250 mmHg (= 33,3 kPa)	>=33kPa	>= 37,5 kPa
Heart rate	>=100/min	>=100/min	>= 80/min
Cardiac index	<2,1 l/min/m ²		
		>0,5	
Total O ₂ extraction		Based on SvO ₂	>0,46
		Based on ScvO ₂	>0,51
CVP	>= 17 mmHg	>= 17 mmHg	> 11 mmHg
Arterial bicarbonate	< 21 mmol/l		
Base excess		< -4 mmol/l	< -5 mmol/l

CPB = cardio pulmonary bypass, IABP = intra-aortic ballon pump, A-a O₂ gradient=alveolar-arterial oxygen gradient, CVP = central venous pressure, SvO₂ = mixed venous saturation, ScvO₂ = central venous saturation.

While still using the same variables, the new model is slightly more complex to calculate (Table 14).

Table 14. Final model (Multiple cut-offs and log-transformed score).

Variable	Measurement interval	Final Coefficients
Age, years	>65 to 70	0.38
	>70 to 75	0.76
	>75 to 80	1.06
	>80	1.25
Albumin, g/l	35 to <40	0.39
	30 to <35	0.81
	<30	1.17
IABP/VAD	Yes	1.76
BE, mmol/l	< -7	0.99
	-7 to <-5	0.5
	> 4	0.89
CVP, mmHg	>16	0.57
ECC-time , min	>60 to 100	0.49
	>100 to 120	0.76
	>120 to 150	1.02
	>150 to 250	1.36
	>250 to 300	1.80
	>300	2.07
Heart rate, bpm	>90	0.59
Creatinine, umol/l	>90 to 110	0.30
	>110 to 150	0.62
	>150 to 200	0.91
	>200	1.68
BSA , m ²	1.7 to 2.0	0.36
	<1.7	0.61
Aa-gradient, kPa	>30 to 40	0.33
	>40 to 50	0.47
	>50	0.80
Oxygen extraction, %	>40 to 50	0.33
	>50	0.62

BE = base excess, CVP = central venous pressure. ECC = extracorporeal circulation. FiO₂ = inspired oxygen fraction. BSA = body surface area, Aa-gradient = alveolar-arterial oxygen gradient, IABP/VAD = intra-aortic balloon pump or ventricular assist device. Oxygen extraction was calculated from mixed venous or central venous saturation.

Discrimination of the new model in the validation dataset was good (aROC of 0.86) compared to that described using general intensive care risk adjustment models in cardiac surgical intensive care [33,34]. Using measurements available on arrival to the ICU is potentially interesting to calculate a change in risk from preoperative to postoperative caused by circumstances during surgery [72].

One potential problem with the model is that differences in CVP, oxygen extraction, ECC-time, IABP/heart assist and heart rate could be the effect of different routines. While the values of these variables do describe an increased risk on admission to the ICU, it may not be the values of the variables themselves that increase the risk but they select a group of patients having higher risk. Different local management routines may affect the values of these variables whereby the values of the variables may not necessarily reflect differences in risk. The problem that variables may be affected by local routines has similarities with lead-time bias [31] and may occur with other models used for risk prediction after cardiac surgery [36–38].

Calibration was not perfect in the validation dataset (SMR 0.86), but the slope of Cox's calibration regression remained very good in the validation dataset (1.01) making overfitting unlikely. The low Cox's calibration intercept (-0.16) in the validation dataset is something the model has in common with the ARCTIC model [37]. SMR fell continuously over the period of the study giving the impression of an improvement in care (Figure 22).

The fall in SMR, however, was not caused by a decrease in observed mortality but rather an increase in predicted mortality, mainly the result of increased ECC-time, lower preoperative albumin values and thirdly that more patients had an abnormal base excess. The increase in ECC-time was mainly due to longer CABG procedures, the reason may be that the number of percutaneous coronary interventions has increased dramatically and CABG is now reserved for more complicated cases. The question is, does a long ECC-time increase risk or is this merely a proxy for more seriously ill patients, more complicated surgery or increased aortic cross clamp time? A possible explanation for the lower preoperative serum albumin values may be that a larger proportion of patients who have had surgery during recent years have had a long preoperative hospital stay or more chronic diseases.

Figure 22. Fall in standardised mortality ratio over the period of the study. SMR=Standardised mortality ratio. Confidence intervals calculated with bootstrapping using 1000 replications.

A shift in predicted mortality rate caused by changes in case-mix may be counteracted by periodic recalibration of the prediction model to keep it up-to date. However, periodic recalibration makes it difficult to compare risk-adjusted mortality rates over time. Furthermore, the low number of deaths after cardiothoracic surgery makes comparison over short periods of time more difficult because of wide confidence limits. While certainly better than raw mortality rates, this reduces the value of risk adjusted mortality rates when following the quality of cardiothoracic intensive care. Analysing trends in risk adjusted mortality over time and why they have occurred is, however, a way of demonstrating improvement (or worsening) in quality of care. Even if a risk adjustment model performs well, differences in risk adjusted mortality rates should be analysed carefully rather than trusted as accurately indicating changes in quality of medical care.

Quality of SIR data

There are advantages and disadvantages with register-based research. The amount of data is large, but the quality of data is generally less dependable than in smaller studies. We therefore need to consider criteria for the quality of data.

Representative of the country

This is fulfilled since all general ICUs, all burn ICUs, all paediatric ICUs, 80 % of the neuro ICUs and 88 % of the cardiothoracic ICUs are members. The coverage of SIR has evolved over time as shown in Figure 3. The good coverage in recent years is positive, but the increase in coverage may be a problem for the risk adjustment models used if the units have important differences. It seems that the SAPS3 model perform wells over time (Figure 21).

For cardiothoracic ICUs the increase in coverage has been proportionately greater and the SMR has decreased, predominantly due to an increase in ECC-time which, as discussed above, may only be a proxy for other more important factors.

The completeness of recruitment

Units participating in SIR recruit consecutive admissions and very few patients refuse registration of their data.

Variables included

SIR covers all important variables. One variable not mentioned that would be valuable to include is a measure of dependency before admission.

Completeness of variables

As described above, there are data missing in SIR regarding physiological parameters in the SAPS3 model and the cardiothoracic intensive care model. The effect of this loss of data is less than expected as described. Administrative data and reason for admission are mandatory and complete. The amount of missing data in a population affects the calibration of a model. The pattern of missing data also affects model performance. Missing data may have different causes and patterns in different countries, especially when there is lack of resources or are financial incentives. This decreases the generalisability of research on missing data. When comparing ICUs internationally, the effect of missing data on validity is difficult to evaluate since the amount of missing data in studies on SAPS3

in other countries is seldom known; it was not described in the original article [24,73].

Collection of raw data

All continuous data are collected in raw format.

Explicit definitions/rules for variables

SIR has guidelines with definitions of all variables used. To be valid in international comparisons, definitions of variables in the register must follow international consensus. SIR has used the original definitions whenever an explicit definition exists. There should be explicit rules defining how the variables are recorded, for example timing of physiological measurements. This is described in the guidelines presented by SIR.

Reliability of coding

Coding of conditions and interventions should be reliable. The reliability of data collection in SIR is not universally tested and this can be considered one of the weakest areas of SIR. The inter-rater reliability of SAPS3 coding has previously been shown to be adequate in Norway [74].

Independence of observations of primary outcome

This is fulfilled in SIR since all deaths are automatically imported from the Swedish National Population Register.

Data validation

SIR data are validated as far as being within prespecified limits and not illogical. Should these criteria not be met, the data are returned for validation. External validation of coding systems for assessment of reason for admission to the ICU and for comparison of measurements recorded in SIR including laboratory results and electronic healthcare records, however, has not been carried out.

The SIR database performs well against the criteria for clinical databases defined in the U.K. [44]. Overall, the performance of the SIR database is comparable to that reported for the U.K. Case Mix Programme Database in 2004 [75]. Sweden has the advantage of easily captured long-term mortality.

But SIR lacks an external validation of the data entered by affiliated ICUs. When models are built or calibrated, having poor quality data, de-

creases the generalisability of the model, and if comparing units, results could be biased by differences in quality of data.

Our evaluation of the performance of SAPS3 in Swedish intensive care should be correct, as should be the case for the cardiothoracic intensive care model. But international generalisability of the amount and effect of missing data is low, as is the comparison of 30-day mortality with in-hospital mortality, since in-hospital mortality is severely biased by availability of long-term acute care or hospice care facilities, discharge policies, financial incentives and the number of transfers between hospitals. Poor international generalisability is also true for the customised cardiothoracic intensive care model as treatment customs vary between countries. Comparison of different time-fixed outcome measures should be quite generalisable since these are not biased by hospital discharge practices, but such outcomes may be affected by case-mix variation, including age.

So, what is the significance of risk adjustment in Swedish intensive care?

Risk adjustment systems are better than raw mortality rates for comparing quality of care since they take the status and history of the patients into account. The risk adjustment systems used in adult intensive care in Sweden perform well, despite some missing data.

While this decreases the effect of differences in case-mix and degree of sickness, it cannot be expected to rule them out completely. Can we rule out differences in case-mix when risk adjusted mortality rates differ? Already when SAPS3 was developed it became evident that case-mixes varied in different parts of the world and local customisation of the formula was used to account for this [24]. Even if the risk adjustment system is well calibrated and performs well, the exact values of risk adjusted mortality cannot be relied upon as a measure of the quality of medical care. There are several reasons for this. The measure of mortality may be biased by different discharge routines. In Sweden we have ruled out bias of in-hospital mortality due to different discharge routines by using time-fixed mortality endpoints. There may be differences in how reason for ICU admission is assessed and how and when physiological variables are sampled. There may be differences in the amount of data missing. All in all, risk adjustment is not perfect. It does not take all aspects of the patient into account. For example, the risk adjustment systems currently used in Sweden do not take frailty into account and frailty has been shown to be a predictor of mortality [76]. If the frailty of patients is different between units, the risk adjusted mortality rate may differ even if the estimated mortality rate from the model is the same. The SAPS3 model has

remained stable over time in Sweden (Figure 21), but the SMR of the cardiothoracic intensive care model developed in Sweden has decreased over time (Figure 22). Though this decreasing in SMR may be caused by improved care, it is more likely to have been caused by a different case-mix which the model is unable to adjust for.

The confidence limits of SMR are quite wide for one year and one unit (Figure 23). And even wider for the cardiothoracic units with fewer deaths (Figure 24). This can be addressed by using longer time series when making comparisons, but this decreases the temporal resolution. There is also the Yule-Simpson's paradox [77] that describes how high-risk admissions have a greater impact on the SMR than low-risk admissions. The proportion of low-risk/high-risk admissions affects the effect that change in SMR within groups has on the overall measure. For example, if the SMR is lower for low-risk admissions and higher for high-risk admissions, this will have a different impact on the overall SMRs depending on the proportions of low-risk and high-risk admissions. The units with large proportions of high-risk admissions will have an increased SMR while the units with big proportions low-risk admissions will have a decreased SMR.

Risk adjustment models can be used to describe how sick patients are, and can act as an alarm bell when SMR increases. It can be used as one of several quality measurements and analysed together with others, but one must be sceptical of precise figures. A good way of using risk adjustment to improve care is to follow trends. An increase in SMR over time may be revealed and action taken. The same model can then be used to follow any improvement in that trend.

The use of a customised SAPS3 model inevitably means that we cannot compare intensive care quality between countries, a fact that was already evident when SAPS3 was developed and customised versions for various parts of the world were published. The question is, however, is it possible to compare intensive care internationally at all? There are several problems involved; the spectrum of diseases differs in various parts of the world, and intensive care may be defined in different ways. Depending on resources such as the number of beds and by availability of long-term acute care or hospice care facilities, patients who are obvious candidates for intensive care may not be admitted to the ICU. There may be financial incentives for an ICU to produce good mortality- or standardised mortality rates, and in such a situation some very sick patients may not be welcome to the ICU.

Risk adjusted mortality does not say anything definite about quality of care, what case-mix there is and how complete registrations are. But if the risk adjusted mortality is good in a unit, then the quality of care and completion of data is sufficient for the disease spectrum/case-mix that exists

on that unit. Risk adjusted mortality should be seen more as a warning bell than an absolute quality variable.

If linked to financial incentives, calibration of the risk adjusted mortality rate may be destroyed by bias in the registration of data. Furthermore, if linked to financial incentives, risk adjusted or raw mortality rates may be a hazardous measure as patients may be denied intensive care in order to keep figures good.

For these reasons risk adjusted mortality rates are difficult to compare internationally. Studies comparing different risk adjustment systems reveal that apparent differences in quality of care depend on the risk adjustment models used and generally conclude that this is caused by one model being more appropriate [78,79]. But this may just as well be caused by the fact that different risk adjustment systems include different risk-factors. When finding differences in risk adjusted mortality, the first measure should be an analysis of the cause of the differences.

Figure 23. The confidence intervals of SMR are wide, one-year SMRs and confidence intervals for general Swedish ICUs are shown. Most of them overlap.

Figure 24. The confidence intervals of SMR are wide when the number of deaths is low. One-year SMRs and confidence intervals for cardiothoracic Swedish ICUs are shown. The confidence intervals are calculated with bootstrapping using 10,000 samples. All confidence intervals overlap.

CONCLUSIONS

The overall aim of the thesis was to study how well risk adjustment models work in Swedish intensive care and what their importance is in the assessment of intensive care results.

We can conclude that the customised version of SAPS3 works well for general intensive care in Sweden.

A Swedish recalibration was needed to achieve good performance of the SAPS3 model. With the original (general) SAPS3 formula the discrimination is good while the calibration is bad. With the recalibrated model both the discrimination and the calibration are generally good. There could certainly be improvements such as adjustment for high age.

Missing variables are common when using SAPS3 in Sweden; 40.9 % of registered admissions miss at least one variable. This loss does not appear totally at random but is linked to low-risk admissions. The model performance is not as badly affected as expected, but the standardised mortality ratio increases when data are missing.

In Sweden the 30-day mortality is slightly higher (17 %) compared to in-hospital mortality (14 %). It is less affected by transfer between units.

We could show that customised versions of SAPS3 works as well for predicting 30-day mortality as for in-hospital mortality.

Customisations of the SAPS3 outcome prediction model perform well with 30-, 90- and 180-day mortality as outcome. Which outcome is preferred for analysis depends on the purpose. For evaluation of care from the patient's perspective, a longer follow-up may be more interesting. For benchmarking of intensive care, 30-day mortality is probably more interesting since it is more influenced by the quality of intensive care than a longer follow-up and few patients remain on the ICU after 30 days.

A model based on data similar to the Higgins-ICU dataset performs well in cardiothoracic intensive care, but the standardised mortality ratio decreases with time. This is not caused by a fall in observed mortality rate but by an increase in expected mortality, mostly based on increased ECC-time.

While certainly better than simply comparing raw mortality rates, the most important insight from this thesis is that comparison of standardised mortality ratios cannot be trusted as indicating differences in quality

of medical care. Before drawing any conclusions, differences revealed should be further analysed.

ACKNOWLEDGEMENTS

I would like to express my gratitude to everyone who has supported me and contributed to the work of this thesis. I specially would like to thank:

Sten Walther, my principal supervisor and colleague. For getting me through this by guiding and inspiration. For many interesting discussions. For your knowledge, your logical reasoning and ability to pick out the most important content in a text.

Folke Sjöberg, my co-supervisor. For your enthusiasm, your great knowledge about research and important discussions on research issues from a broader perspective.

Mats Fredrikson, my co-supervisor. For statistical advice when needed.

My other co-authors: Kristina Rydenfelt, Ulf Strömberg, Carolina Samuelsson, Andrew A. Kramer, Thomas Nolin, Göran Karlström, Caroline Mårdh, Wolfgang Freter and Johan Sellgren. For your important contributions to this work.

Peter Cox, my colleague. For enthusiastic, thorough and extremely fast help with the English language.

Robert Svensson, my colleague and current chief. For enthusiastic participation in investigation of the changed risk adjusted mortality rate at our department, the origin of this thesis.

Colleagues at the department of anesthesiology and intensive care at Vrinnevi hospital and the department of thoracic and vascular surgery at the university hospital in Linköping. For making it a pleasure to go to work.

Ingegerd and Bengt, my parents. For their introduction to the joy of investigation of how things work. For placing a high value on education. And for striving for fairness and environment rather than materialistic goals.

Anna and Elin, my sisters. For support and enthusiasm. For interesting discussions with different perspectives. And for helping me getting out into the nature.

My family, Carina, Isabelle and Erik. Because you are everything. And for keeping me focused on important things in life.

REFERENCES

1. Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. Historical perspective on hand hygiene in health care [Internet]. World Health Organization; 2009 [cited 2018 Mar 1]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK144018/>
2. Chun J, Bafford AC. History and Background of Quality Measurement. Clin Colon Rectal Surg. 2014;27:5–9.
3. Power GS, Harrison DA. Why try to predict ICU outcomes? Curr Opin Crit Care. 2014;20:544–9.
4. Riktlinjetråd | Svensk Förening för Anestesi och Intensivvård [Internet]. [cited 2018 Feb 20]. Available from: <https://sfai.se/riktlinje/riktlinjetrad/>
5. Intensivvård i Vården i siffror [Internet]. [cited 2018 Feb 20]. Available from: <https://vardenisiffror.se/category/sjukdomstillstand/intensivvard>
6. Befolkningsstatistik [Internet]. Stat. Cent. [cited 2018 Mar 27]. Available from: <http://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/befolkningens-sammansattning/befolkningsstatistik/>
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;38:1647–53.
8. Strand K, Walther SM, Reinikainen M, Ala-Kokko T, Nolin T, Martner J, et al. Variations in the length of stay of intensive care unit nonsurvivors in three Scandinavian countries. Crit Care. 2010;14:R175.
9. Fitz-Henry J. The ASA classification and peri-operative risk. Ann R Coll Surg Engl. 2011;93:185–7.
10. Lane-Fall MB, Neuman MD. Outcomes measures and risk adjustment. Int Anesthesiol Clin [Internet]. 2013 [cited 2018 Feb 23];51. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3816010/>
11. 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 Engl. 2014;14:376–9.
12. Dripps RD, Lamont A, Eckenhoﬀ JE. The role of anesthesia in surgical mortality. JAMA. 1961;178:261–6.

13. Zimmerman JE, Kramer AA. A history of outcome prediction in the ICU. *Curr Opin Crit Care*. 2014;20:550–6.
14. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med*. 1981;9:591–7.
15. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med*. 1985;13:818–29.
16. 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;12:975–7.
17. Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. *Crit Care Med*. 1985;13:519–25.
18. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*. 1991;100:1619–36.
19. 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;270:2957–63.
20. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA*. 1993;270:2478–86.
21. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34:1297–310.
22. Cook SF, Visscher WA, Hobbs CL, Williams RL, Project IMPACT Clinical Implementation Committee. Project IMPACT: results from a pilot validity study of a new observational database. *Crit Care Med*. 2002;30:2765–70.
23. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med*. 2007;35:827–35.
24. Moreno RP, Metnitz PGH, 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;31:1345–1355.

25. Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. *Crit Care Med*. 2007;35:1091–8.
26. Harrison DA, Ferrando-Vivas P, Shahin J, Rowan KM. Ensuring comparisons of health-care providers are fair: development and validation of risk prediction models for critically ill patients [Internet]. Southampton (UK): NIHR Journals Library; 2015 [cited 2017 Nov 15]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK321613/>
27. Graham PL, Cook DA. Prediction of risk of death using 30-day outcome: a practical end point for quality auditing in intensive care. *Chest*. 2004;125:1458–66.
28. Brinkman S, Abu-Hanna A, de Jonge E, de Keizer NF. Prediction of long-term mortality in ICU patients: model validation and assessing the effect of using in-hospital versus long-term mortality on benchmarking. *Intensive Care Med*. 2013;39:1925–31.
29. Glance LG, Szalados JE. Benchmarking in critical care: the road ahead. *Chest*. 2002;121:326–8.
30. Boyd O, Grounds RM. Physiological scoring systems and audit. *Lancet*. 1993;341:1573–1574.
31. Tunnell RD, Millar BW, Smith GB. The effect of lead time bias on severity of illness scoring, mortality prediction and standardised mortality ratio in intensive care—a pilot study. *Anaesthesia*. 1998;53:1045–1053.
32. Ferrando-Vivas P, Jones A, Rowan KM, Harrison DA. Development and validation of the new ICNARC model for prediction of acute hospital mortality in adult critical care. *J Crit Care*. 2017;38:335–9.
33. Doerr F, Badreldin AMA, Can F, Bayer O, Wahlers T, Hekmat K. SAPS 3 is not superior to SAPS 2 in cardiac surgery patients. *Scand Cardiovasc J SCJ*. 2014;48:111–9.
34. Martínez-Alario J, Tuesta ID, Plasencia E, Santana M, Mora ML. Mortality prediction in cardiac surgery patients: comparative performance of Parsonnet and general severity systems. *Circulation*. 1999;99:2378–82.
35. Higgins TL, Estafanous FG, Loop FD, Beck GJ, Lee JC, Starr NJ, et al. ICU admission score for predicting morbidity and mortality risk after coronary artery bypass grafting. *Ann Thorac Surg*. 1997;64:1050–8.
36. Hekmat K, Kroener A, Stuetzer H, Schwinger RHG, Kampe S, Bennink GBWE, et al. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. *Ann Thorac Surg*. 2005;79:1555–62.

37. Shahin J, Ferrando-Vivas P, Power GS, Biswas S, Webb ST, Rowan KM, et al. The Assessment of Risk in Cardiothoracic Intensive Care (ARCtIC): prediction of hospital mortality after admission to cardiothoracic critical care. *Anaesthesia*. 2016;71:1410–6.
38. Tamayo E, Fierro I, Bustamante-Munguira J, Heredia-Rodríguez M, Jorge-Monjas P, Maroto L, et al. Development of the Post Cardiac Surgery (POCAS) prognostic score. *Crit Care Lond Engl*. 2013;17:R209.
39. Lamarche Y, Elmi-Sarabi M, Ding L, Abel JG, Sirounis D, Denault AY. A score to estimate 30-day mortality after intensive care admission after cardiac surgery. *J Thorac Cardiovasc Surg*. 2017;153:1118–1125.e4.
40. Slater A, Shann F, Pearson G, Paediatric Index of Mortality (PIM) Study Group. PIM2: a revised version of the Paediatric Index of Mortality. *Intensive Care Med*. 2003;29:278–85.
41. 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 paediatric intensive care*. *Pediatr Crit Care Med J Soc Crit Care Med World Fed Pediatr Intensive Crit Care Soc*. 2013;14:673–81.
42. Start - SIR-Svenska Intensivvardsregistret [Internet]. [cited 2018 Feb 27]. Available from: <http://www.icuregswe.org/>
43. Emilsson L, Lindahl B, Koster M, Lambe M, Ludvigsson JF. Review of 103 Swedish Healthcare Quality Registries. *J Intern Med*. 2015;277.
44. Black N, Payne M. Improving the use of clinical databases. *BMJ*. 2002;324:1194.
45. Ludvigsson JF, Andersson E, Ekbom A, Feychting M, Kim J-L, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. *BMC Public Health*. 2011;11:450.
46. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A. The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *Eur J Epidemiol*. 2009;24:659–667.
47. 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;21:128–138.
48. Nassar AP, Malbouisson LMS, Moreno R. Evaluation of simplified acute physiology score 3 performance: a systematic review of external validation studies. *Crit Care*. 2014;18:R117.
49. Cox D. Two further applications of a model for binary regression. *Biometrika*. 1958;45:562–5.

50. 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:e16110.
51. Poole D, Carrara G, Bertolini G. Intensive care medicine in 2050: statistical tools for development of prognostic models (why clinicians should not be ignored). *Intensive Care Med*. 2017;
52. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78:1–3.
53. Breslow NE, Day NE. Statistical methods in cancer research. Volume II--The design and analysis of cohort studies. IARC Sci Publ. 1987;
54. Dover DC, Schopflocker DP. Using funnel plots in public health surveillance. *Popul Health Metr*. 2011;9:58.
55. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7:147–77.
56. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J Am Stat Assoc*. 1988;83:1198–202.
57. Little R. Missing-Data Adjustments in Large Surveys. *J Bus Econ Stat*. 1988;6:287–96.
58. Rubin DB, editor. Multiple Imputation for Nonresponse in Surveys [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1987 [cited 2018 Jun 22]. Available from: <http://doi.wiley.com/10.1002/9780470316696>
59. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med*. 1989;8:551–61.
60. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58:267–88.
61. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 1970;12:55–67.
62. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67:301–20.
63. McNEMAR Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12:153–7.
64. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat Med*. 1998;17:2635–50.

65. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539–58.
66. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327:557–60.
67. Buuren S van, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. *J Stat Softw* [Internet]. 2011 [cited 2017 Jun 29];45. Available from: <http://www.jstatsoft.org/v45/i03/>
68. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32–5.
69. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. *J Am Stat Assoc*. 1979;74:829–36.
70. Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. *J Interdiscip Math*. 2010;13:253–67.
71. Afessa B, Keegan MT, Gajic O, Hubmayr RD, Peters SG. The influence of missing components of the Acute Physiology Score of APACHE III on the measurement of ICU performance. *Intensive Care Med*. 2005;31:1537–43.
72. Coulson TG, Bailey M, Reid CM, Tran L, Mullany DV, Smith JA, et al. Acute Risk Change for Cardiothoracic Admissions to Intensive Care (ARCTIC index): a new measure of quality in cardiac surgery. *J Thorac Cardiovasc Surg*. 2014;148:3076-3081.e1.
73. Metnitz PGH, 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;31:1336–1344.
74. Strand K, Strand LI, Flaatten H. The interrater reliability of SAPS II and SAPS 3. *Intensive Care Med*. 2010;36:850–3.
75. Harrison DA, Brady AR, Rowan K. Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit & Research Centre Case Mix Programme Database. *Crit Care Lond Engl*. 2004;8:R99-111.
76. Flaatten H, De Lange DW, Morandi A, Andersen FH, Artigas A, Bertolini G, et al. The impact of frailty on ICU and 30-day mortality and the level of care in very elderly patients (≥ 80 years). *Intensive Care Med*. 2017;43:1820–8.
77. Simpson EH. The Interpretation of Interaction in Contingency Tables. *J R Stat Soc Ser B Methodol*. 1951;13:238–41.

78. Kramer AA, Higgins TL, Zimmerman JE. Comparing observed and predicted mortality among ICUs using different prognostic systems: why do performance assessments differ? *Crit Care Med.* 2015;43:261–9.
79. Kramer AA, Higgins TL, Zimmerman JE. Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: implications for national benchmarking*. *Crit Care Med.* 2014;42:544–53.

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-151308>