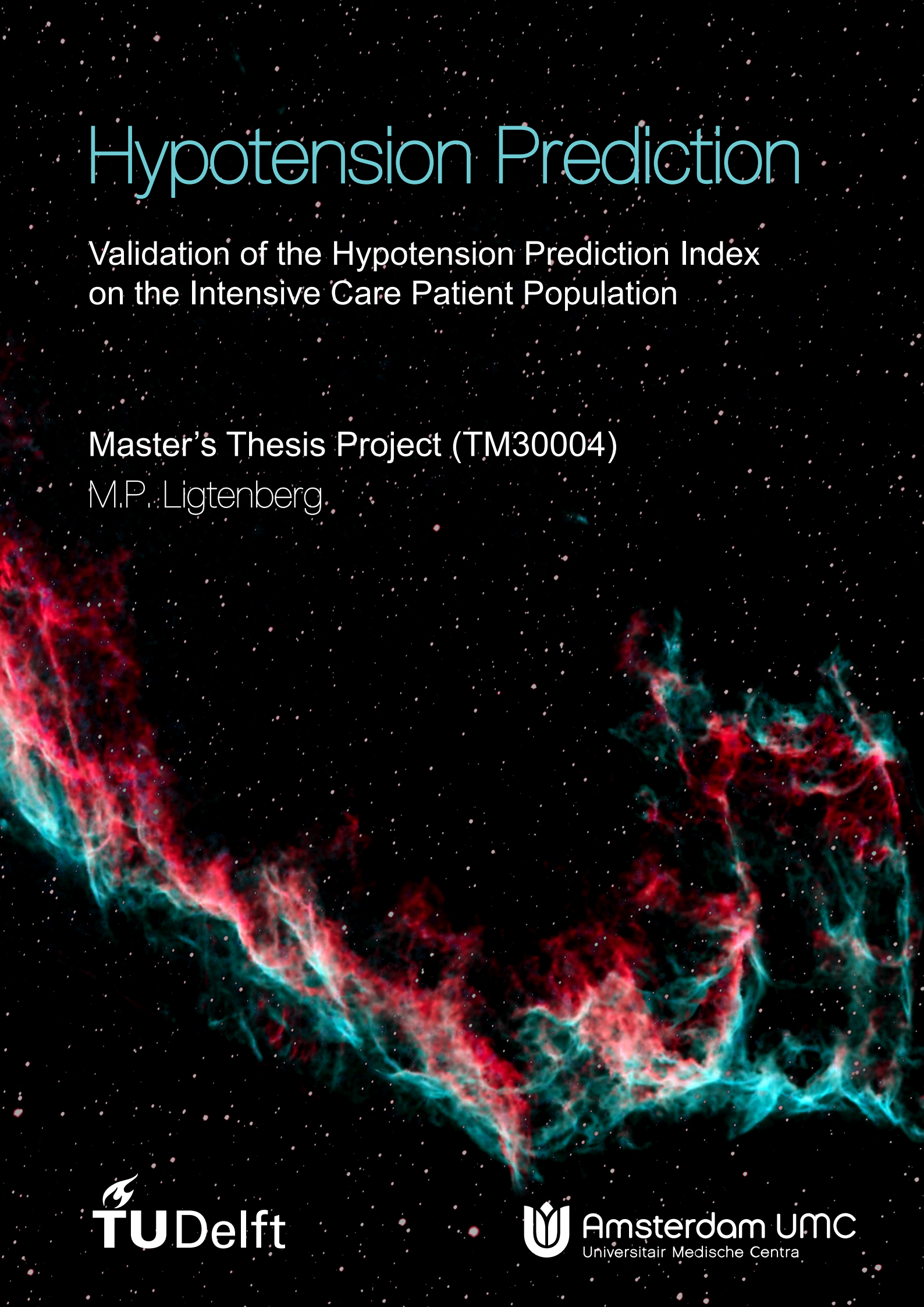
Hypotension Prediction

Validation of the Hypotension Prediction Index on the Intensive Care Patient Population

Master's Thesis Project (TM30004)

M.P. Ligtenberg



TU Delft

Amsterdam UMC
Universitair Medische Centra

Hypotension Prediction

Validation of the Hypotension Prediction Index on the Intensive Care Patient Population

by

M.P. Ligtenberg

04-11-2021

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in Technical Medicine
Leiden University | Delft University of Technology | Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)
Dept. of Intensive Care Medicine, Amsterdam UMC, Location AMC
April to November 2021
Student no.: 4390075

Thesis assessment committee members

prof. A.P.J. Vlaar (Alexander), MD, PhD, MBA	Medical supervisor
E. Demirović (Emir), PhD	Technical supervisor
J.P. van der Ster (Björn), PhD	Daily supervisor
prof. E. de Jonge (Evert), MD, PhD	Chair

Institutions
Delft University of Technology
Amsterdam UMC, Location AMC

TU Delft

Amsterdam UMC
Universitair Medische Centra

Preface

My first visit to the intensive care was at the age of eleven, when my beloved grandmother was fighting for her life. There are two things I can recall: she being too weak to lift her arm, but also the impression of monitors and technology around her. More then a decade later, I can say she sparked my interest for Technical Medicine and I proudly present her my Master's Thesis.

The master thesis project is the pinnacle of the master programme of Technical Medicine. This interdisciplinary joint-degree Master programme is offered by Delft University of Technology, together with Leiden University, Erasmus University Rotterdam and their academic medical centers. Nonetheless, I conducted my Master thesis project at the department of Intensive Care Medicine of the Amsterdam University Medical Centre, location AMC, within the Dynamo hemodynamics research group.

First and foremost, I would like to thank Björn van der Ster, my daily supervisor, for our talks, honest feedback and direct supervision of my project. Our Monday morning talks with coffee formed the foundation of this project. I appreciate your friendliness, tolerance and sharp feedback.

I thank my medical supervisor Alexander Vlaar and fellow principal investigator Denise Veelo for the facilitation of the project. Your invitation to the research group felt warm since our first meeting. Hopefully, you will continue to focus on the clinical usefulness of new technology, the clinical evaluation and maybe even the large-scale implementation of the Hypotension Prediction Index module.

Also, I want to thank Emir Demorović, my technical supervisor from the faculty of Engineering, Mathematics and Computer Science of Delft University of Technology. Thank you for taking the leap into the medical domain and providing technical guidance on the project. You really helped me to further develop my academic writing skills.

As the last of colleagues, I want to show my appreciation to my direct colleagues roommates Jaap, Santino and Tineke, but also to all other researchers of the department, for the good times we had together at the AMC.

Pap en mam, Vince, thank you for you unconditional support throughout every situation. In the end, you are the driven piles of what I have built so far. Pap, this print wouldn't have been eye candy without you.

Tessel, thanks for dealing with my daily challenges. You gave me the support and distraction I needed last year.

M.P. Ligtenberg
Rotterdam, October 2021

Summary

A low blood pressure (hypotension, as mean arterial pressure < 65 mmHg) in patients on the intensive care unit (ICU) is associated with adverse outcomes and death. Pro-active treatment with the use of a predictive alarm could improve patient outcome. The Hypotension Prediction Index (HPI) is a machine-learning algorithm that uses arterial blood pressure waveforms to calculate the probability of impending hypotension. Prior to clinical implementation, the HPI model needs to be validated. In previously published validation protocols of the HPI, possible sources of bias were identified. Therefore, the primary objectives were to assess HPI performance on the ICU population by using a clinically relevant validation protocol and to evaluate the differences between previous protocols. Secondary objectives included evaluation of subgroup performance and the effect of individual validation protocol settings on the subsequent results.

The three applied validation protocols used conceptually different methods to classify predictions (e.g. as a true or false alarm). The primary forward sliding window (FSW) protocol labels each single prediction based on hypotension occurrence thereafter. The forward tumbling window (FTW) groups predictions in non-overlapping time windows to only classify the window based on hypotension occurrence thereafter. The backward (BW) protocol labels the onsets of hypotension based on alarm occurrence at *'t'* minutes prior to onset, hence 'backward' in time. Identical metrics were used to quantify performance. For secondary analyses the FSW protocol was used.

Performance for the FSW protocol was reduced compared to the FTW and BW protocol. Using the FDA-approved alarm settings for FSW, FTW and BW (t=10min) protocols, sensitivity was 0.59, 1.00 and 0.83, respectively. Positive predictive values were 0.41. 0.83 and 1.00, respectively. For the FSW protocol the median [IQR] time-to-hypotension was 3.3 [1.0 to 7.3] min, for an alarm defined as the last of consecutive alarms prior to hypotension. Reducing the minimal mean arterial pressure in the definition of non-hypotension from 75 to 65 mmHg reduced the area under the precision recall curve from 0.82 to 0.31. Omission of a washout period of 30 min increased the positive predictive value from 0.41 to 0.64.

This thesis demonstrates the importance of validation methodology and the generalizability of the Hypotension Prediction Index to the intensive care unit population. The definition of non-hypotension and a washout period strongly influenced the results. Overall, the results demonstrate the ability of the HPI to predict hemodynamic instability in ICU patients. Therefore, validation results support the introduction of the HPI to the ICU for clinical use. However, the optimal alarm threshold and clinical benefit remain to be evaluated in future clinical studies.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ABP	arterial blood pressure
AUC	area under the curve
AUCPR	area under the PR curve
AUROC	area under the ROC curve
BW	backward (validation protocol)
DBP	diastolic blood pressure
FN	false negative
FP	false positive
FPR	fasle postive rate: 1 - specificity
FSW	forward sliding window (validation protocol)
FTW	forward tumbling window (validation protocol)
HPI	hypotension prediction index
ICU	intensive care unit
MAP	mean arterial blood pressure
mmHg	millimetres of mercury, unit of pressure, 100 mmHg = 13,3 kPa
NPV	negative predictive value
PPV	positive predictive value
PR curve	precision-recall curve
ROC curve	receiver operator characteristic curve
Recall	sensitivity
SBP	systolic blood pressure
SE	sensitivity
SP	specificity
TN	true negative
TP	true positive
TPR	true positive rate, see sensitivity

Equations

$$\text{Blood Flow } = \frac{\text{Blood Pressure}}{\text{Vascular Resistance}} \tag{1}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$PPV = Precision = \frac{TP}{TP + FP} \tag{4}$$

$$NPV = \frac{TN}{TN + FN} \tag{5}$$

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall} = \frac{PPV \times Sensitivity}{PPV + Sensitivity} \tag{6}$$

Table 2: Contingency table

		Observation Hypotension	Non-hypotension		
HPI	Alarm	True positive (TP)	False positive (FP)	↔	PPV = TP/(TP+FP)
	No-alarm	False Negative (FN)	True Negative (TN)	↔	NPV = TN/(TN+FN)
		↕	↕		
		Se = TP/(TP+FN)	Sp = TN/(TN+FP)		

Abbreviations: *Se*, sensitivity; *Sp*, specificity; *PPV*, positive predictive value; *NPV*, negative predictive value

List of Figures

List of Tables

Introduction

The Veil Nebula is a cloud of heated and ionized gas. It is a galactic supernova remnant from an explosion 10,000-20,000 years ago. The red and blue colours of the nebula are a metaphor for the oxygen-rich arterial blood and the oxygen-poor venous blood. Just like the Hypotension Prediction Index model, this image was constructed using features invisible to the naked eye and many hours of data.

Introduction

Organ function fundamentally relies on adequate oxygen supply for its cellular metabolism. Maintenance or restoration of oxygen supply is therefore important to preserve organ function, especially in critically ill patients on the Intensive Care Unit (ICU) as oxygen consumption is increased.[38] Shock is the clinical state of circulatory failure that leads to an inadequate blood supply to vital organs. This results in hampered oxygen delivery and an oxygen deficit in tissue. Severity and duration of tissue hypoxia (i.e. shortage of oxygen in tissue) correlates with complications, organ failure and death.[46]

On a cellular level, hypoxia is damaging in multiple ways. Hypoxia causes energy metabolism in mitochondria to switch from aerobic to anaerobic breakdown of glucose. This produces an excess amount of lactate, hydrogen ions and inorganic phosphates. Primary energy molecule levels in cells, Adenine triphosphate (ATP), decrease because of diminished production and continued consumption. Low ATP levels hamper production of proteins, which imperils mitochondrial function and eventually leads to cell death and end-organ failure.[1,18]

Unfortunately, options to identify an oxygen deficit in tissue of patients are limited in clinical practice. Clinicians have to rely on clinical presentation of the patient and surrogate (bio) markers to determine adequate tissue oxygenation.[38]

Tissue oxygenation is, amid other factors, generally determined by blood flow. However, blood flow cannot be measured easily and is mostly measured invasively. According to the laws of science, flow rate is proportional to driving pressure and inversely proportional to the resistance. So, blood flow depends on blood pressure and vascular resistance. Therefore, one of the indirect markers that indicate circulatory stability is arterial blood pressure.[38]

Hypotension is the state of having a low arterial blood pressure and an indicator for an oxygen deficit in tissue. Hypotension is associated with a higher occurrence of acute kidney injury and a higher mortality rate during noncardiac surgery and patients with sepsis on the intensive care unit (ICU).[3,14,31,33,52] Prevention of hypotension is therefore included in treatment guidelines for septic patients.[29,45] However, incidence of hypotension in the intensive care unit reportedly still ranges from 23% to 72%.[16,33,47,53]. Reducing hypotensive event duration or severity could lead to improved clinical outcome of patients.[14,43]

Current treatment of hypotension is mainly reactive. Possible hemodynamic interventions include administration of fluids, medication that predominantly increases peripheral vascular resistance and medication that stimulates the contractility of the heart.[1,14,38] Hypotensive episodes can occur unexpectedly and suddenly. As hemodynamic interventions are applied reactively, intervention is only initiated once the patient has already entered a hypotensive episode. A limitation of reactive treatment is the delayed effect. Once the intervention is prepared and comes into effect, the blood pressure could have decreased even further.

Alarms on impending hypotension could enable pro-active treatment. Multiple predictive computer models have been designed to alarm for impending hypotension.[8,9,19,25,39] Such an alarm would both enable clinicians to pro-actively start treatment and increase time for clinicians to prepare hemodynamic intervention. Timely intervention by use of a predictive model could reduce hypotension occurrence and severity in intensive care patients. This effect has already been demonstrated in surgical patients. During noncardiac surgery, pro-active treatment with the use of the Hypotension Prediction Index (Edwards Lifesciences, Irvine, CA, USA) resulted in less intraoperative hypotension, without increasing the amount of medication administered.[59]

The Hypotension Prediction Index (HPI) algorithm is a machine-learning derived algorithm that alarms for impending hemodynamic instability.[19] It is designed on arterial blood pressure wave form data from both surgical and intensive care patients. A logistic regression model was used to calculate the predictive HPI value, which was scaled between 0 and 100. More details are explained in Appendix A. A higher HPI value should be interpreted as a greater calculated probability of impending hypotension. A threshold value for HPI is used to dichotomize the continuous warning scale into 'alarm' and 'no-alarm' of impending hypotension. In clinical practice, an HPI value that exceeds the threshold value triggers a bed-side alarm. This suggests initiation of a hemodynamic intervention to prevent hypoperfusion of critical organs.[19]

Prior to any clinical implementation, predictive computer models require thorough retrospective quality assessment. During validation of a predictive model, its generalizability is assessed, i.e. the degree in which model predictions correspond with the actual outcomes. Correct validation methodology is of paramount importance in evaluating the potentially added clinical value of the model, as well as the potentially induced harm due to incorrect predictions.[48]

Designing the validation method is a challenging task as it is subject to the intention of the validation and to the setting in which the predictive model would be used. Previously published validation protocols for hypotension prediction models for intensive care patients showed that the onset of hypotension could be predicted. However, previous validation protocols may show overoptimism in their presented results, when translated to clinical applicability. This was concluded in the literature study conducted prior to this master's thesis (Appendix J). Possible bias of results on algorithm performance was introduced by the following aspects: Firstly, use of performance metrics was often not justified as they could be inflated by class imbalance, i.e. the skewed ratio of hypotension versus non-hypotension occurrence. Secondly, some results of statistical features were presented for clinically unrealistic alarm settings. Lastly, in several statistical analyses, blood pressure datapoints were ignored if they showed near-hypotension values. Removing this 'twilight zone' from analysis creates more contrast between the hypotension and non-hypotension scenarios. Therefore, it becomes easier for the model to give a correct prediction and to achieve better scores on performance metrics. Removing near-hypotension values arguably polarises the results unrealistically.

All aspects mentioned above are ought to contribute to biased results as they do not depict the real clinical situation in which the predictive model may be used. Therefore, generalizability of validation results could be highly impaired. This impairment could lead to erroneous decision making on clinical implementation of the algorithm, with potential harm or untouched benefit to patients as a result.

The primary objectives of this Masters' Thesis were to assess the performance of the HPI on

the intensive care population, by using a clinically relevant validation protocol and to evaluate the differences with previously published protocols. Validation protocols cover data preprocessing, (non-)hypotension classification and a selection of performance metrics to evaluate discriminative performance, calibration and timeliness of the algorithm.

Secondary objectives of this Masters' Thesis included the comparison between subgroups of patients on algorithm performance. Additionally, the effect of multiple settings in validation methodology that could bias results on algorithm performance were evaluated.

Background

Background

Core concepts that form the foundation of this masters' thesis are discussed in this section. This includes elaboration on the definition of hypotension, general elements of a validation protocol and the description of the three validation protocols applied in this thesis.

2.1. Definition of hypotension

No consensus exists on the clinical definition of hypotension.[4] Many different definitions for hypotension exist, with criteria using absolute systolic (SBP), diastolic (DBP) or mean arterial blood pressure (MAP) values and their relative differences to the baseline values of the patient. A commonly used threshold of hypotension is a mean arterial blood pressure of 65 mmHg.[4] This originates form mammal experiments that indicated a lower cerebral self-regulatory threshold at 65 mmHg.[10,11] For blood pressures below this threshold, cerebral tissue was not able to locally reduce vascular resistance to maintain adequate blood flow. However, autoregulatory thresholds differ per organ and per patient.[24,51] For example, patients with a history of chronic hypertension can present with shock symptoms under universally accepted blood pressures.[14]

So, no one-size-fits-all definition of hypotension is applicable. However, a fixed threshold for the definition of hypotension is indeed used for all ICU patients by most ICU clinicians. A recent survey among ICU physicians and nurses showed that the majority used a MAP lower than 65 mmHg to define hypotension.[44]

With hypotension defined as a MAP < 65 mmHg, hypotension is associated with several complications. In general ICU patients, hypotension is found to be a risk factor for the development of acute kidney injury (AKI).[28] In ICU patient with a distributive shock, AKI[2,33,41], myocardial infarction[33] and mortality[2,33,41,58] are associated with hypotension.

2.2. Validation protocol elements

A validation of a hypotensive predictive algorithm is divided into three stages: 1) Data preprocessing, 2) labeling of data according to definitions of (no-)alarms and (non-)hypotension and 3) calculation of performance metrics. Each individual validation protocols will be explained according to these three stages.

2.2.1. Data preprocessing

Data preprocessing is generally defined as the "the collection and manipulation of items of data to produce meaningful information".[13] In this step, data inadequate for analysis is re-

moved. Data points with a bad signal quality or other artefacts deemed as non-physiological are removed or annotated. This step also includes imputation of missing data, i.e. replacing missing values with its estimation.

2.2.2. Data Labelling

Time segments in the data need to be categorized and labelled, to allow calculation of statistical performance metrics of a predictive algorithm. Each label consists of a categorical value pair: the prediction (alarm/no-alarm) and the outcome (hypotension/non-hypotension).

As both values are binary, four types of labels exist. True positives (TPs) are alarms that are followed by hypotension. False positives (FPs) are alarms that are followed by non-hypotension. True negatives (TN) are no-alarms that are followed by non-hypotension. False negatives (FN) are no-alarms that are followed by hypotension.

So, a TP are correctly predicted hypotensive events. A TN is a correctly predicted non-hypotensive event. A FP is a falsely predicted non-hypotensive event, also known as type I errors. A FN is an event that was not predicted and is also known as type II error. All labels can be presented in a contingency table : a tool for model performance assessment that forms the foundation of other performance metrics(Table 2.1).[27]

Prediction values in the form of an HPI value between 0 and 100 are dichotomized by an alarm threshold to an 'alarm' or 'no-alarm'. For example, if a threshold of 50 is used. HPI values greater than 50 will produce an alarm. HPI values lower than or equal to 50 will be a no-alarm. An alarm threshold could form a cut-off value for clinicians to initiate pro-active treatment or not. The performance of an algorithm is often presented for multiple thresholds.

Outcome values are also dichotomized by defining hypotension and non-hypotension. The definitions of hypotension and non-hypotension that the model must predict are a design choice during the development of the predictive algorithm. So, the definitions may vary per algorithm design or published validation protocol.

Table 2.1: Contingency table

		Observation Hypotension	Non-hypotension		
HPI	Alarm	True positive (TP)	False positive (FP)	↔	PPV = TP/(TP+FP)
	No-alarm	False Negative (FN)	True Negative (TN)	↔	NPV = TN/(TN+FN)
		↕	↕		
		Se = TP/(TP+FN)	Sp = TN/(TN+FP)		

Abbreviations: *Se*, sensitivity; *Sp*, specificity; *PPV*, positive predictive value; *NPV*, negative predictive value

2.2.3. Performance metrics

During validation, model quality can be split in two different aspects: discriminative performance and calibration.

Discriminative performance is the ability of the model to separate the different outcomes. Therefore, dichotomized predictions are used in attempt to separate the healthy from the diseased, or the time windows with hypotension from non-hypotension.[50]

Calibration is the agreement between the predicted probability of an event and the observed frequency of events. Therefore, the rate of hypotension occurrence is evaluated for every HPI value from 0 to 100. This is one of the primary requirements to determine clinical usefulness.[48]

Performance metrics quantify the quality of discrimination and calibration of models. Examples of different performance metrics are provided in below in Chapter 2.1.2 of the Literature Study (Appendix J).

Examples of performance metrics

Sensitivity , or 'Recall', is defined as 'the true positive rate' or 'the share of TP of all events. Sensitivity depicts the probability of an alarm given that hypotension will occur shortly.

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{2.1}$$

Specificity is defined as 'the true negative rate' or 'the share of TN of all non-events'. Specificity depicts the probability of a non-alarm given that non-hypotension will occur.

$$Specificity = \frac{TN}{TN + FP} \tag{2.2}$$

Positive Predictive Value (PPV), or 'Precision', is 'the share of TP in all alarms'. PPV depicts the probability that an alarm will be followed by an event, i.e. hypotension.[48] Thus, PPV measures the exactness the positive predictions.[20]

$$PPV = Precision = \frac{TP}{TP + FP} \tag{2.3}$$

Negative Predictive Value (NPV) is 'the share of TN in all non-alarms'. It depicts the probability that a non-alarm will be followed by a non-event.

$$NPV = \frac{TN}{TN + FN} \tag{2.4}$$

Accuracy is the proportion of correct predictions in all predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{\#\, correct\ predictions}{\#\, samples} \tag{2.5}$$

Receiver Operating Characteristic Curve is a frequently used method for analysing discriminative performance. A Receiver Operating Characteristic (ROC) Curve curve displays the performance of the model by plotting the sensitivity against 1-specificity for all threshold values.[5] A point on the ROC curve represents model performance for a single, particular threshold value. Thereby, true positive rate is plotted against false positive rate for the complete range of thresholds. This gives plot gives an overview of model performance.

Thereafter, the ROC curve can also be used to pick the eventual classification threshold, in combination with a cost function. A cost function defines the desired trade-off between sensitivity and specificity, which also forms a line on the ROC plot. The intersection of the ROC curve and cost function indicate the optimal threshold value for that cost function. But also without any cost function, the ROC curve shows valuable information when the operating threshold of the algorithm is yet to be determined.[21]

Via the ROC curve, discriminative ability of a model can be summarised in a single numerical measure: the area under the curve (AUC). The area under the ROC curve (AUROC) is a common technique for evaluating and comparing models on dat asets with varying class distribution.[21] A perfect model has an AUC of 1. A non-informative model has an AUC of 0.5, which is the case when predicting 'heads' of 'tails' when flipping a coin.

Precision Recall Curve displays the Recall (i.e. sensitivity) against the Precision (PPV) for the complete range of thresholds. Thereby, it is similar to the ROC curve. The area under the PR curve (AUCPR) can be used a a summarizing metric.

2.3. Validation protocols used

The three applied validation protocols are described below. Each protocol uses a conceptually different method to label the data. Protocol also used different performance metrics to describe model performance. The protocol of Moghadam *et al.*[37] was originally used on a different hypotension predictive algorithm than the HPI. The original protocol is described below and the necessary adaptations for validation of the HPI are explained in the methods section.

2.3.1. Forward sliding window validation

This continuous forward sliding window (FSW) validation protocol was previously published by Moghadam *et al.*[37]. This research group validated their own hypotension prediction algorithm by using a novel, continuous validation method. This validation methods mimics the clinical scenario of real-time monitoring by performing 'forward' validation of predictions with 1-minute intervals. Data pre-processing consisted of three elements to exclude erroneous data from analysis: 1) the removal of spikes in MAP data in which the variation was > 25% of the baseline in a one-minute window, 2) removal of values outside of the clinical range as listed in Table 2.2, 3) interpolation of physiological values if the missing interval was less than 5 minutes, otherwise that interval was removed.

A hypotensive event was defined as a period of minimally 30 minutes with MAP < 65 mmHg for at least 90% of the time. Non-hypotension was defined as any datapoint with MAP > 75 mmHg.

Table 2.2: Defined clinical range of physiological blood pressure values.

	Minimum	Maximum
Mean arterial blood pressure (mmHg)	30	150
Systolic blood pressure (mmHg)	50	220
Diastolic blood pressure (mmHg)	20	103

Labelling of events was performed using the 'forward' methodology. In forward validation, each individual prediction is annotated first. Thereafter, the label is determined by the occurrence of hypotension in the subsequent period.

The different algorithm evaluated by Moghadam *et al.* [37] is expected to alarm for hypotension 30 minutes in advance. So, an alarm must be followed by the onset of hypotension within the next 30 minutes to be labelled as TP, as illustrated in Figure 2.1 via positive points. Positive points are timestamps in the 30-minute window prior to hypotension. The predictive model is expected to alarm for hypotension on positive points. Regarding predictions made on positive points, an alarm is labelled a TP and a non-alarm is labelled a FN.

Every non-alarm is expected to predict an absence of hypotension for the next 40 minutes. So, every timestamp located at more than 40 minutes before onset of hypotension is pictured as a negative point in figure 3. Regarding predictions made on negative points, a non-alarm is labelled a TN and alarm is labelled a FP.

There is a difference between the time windows that an alarm and a non-alarm cast a prediction on, i.e. 30 minutes and 40 minutes, respectively. These data points are regarded as neutral and were removed for analysis. The neutral points are called *'leading neutral buffer points'*.

Predictions during a hypotensive period were also removed from analysis, as well as predictions made in the 20 minutes after a hypotensive event. In this so-called *'washout'* period of 20 minutes, the physiological state of the patient is deemed to be subject to past hemodynamic interventions. As the hemodynamic state during the washout period is subject to recent interventions, predictions are inaccurate. In addition, the patient would be monitored more closely during the washout period, which reduces the value of a predictive model during that time.

Performance metrics were calculated using a statistically optimal threshold. Metrics consisted of accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Several different statistical methods were used to pick the alarm threshold, of which the results were presented. The alarm threshold with the maximum F1-score was chosen (Equation 2.6). The F1-score is a general, overall performance metric. The Receiver Operating Characteristic (ROC) curve was also used as performance metric. In an ROC curve, the sensitivity is plotted against the sensitivity for all possible alarm thresholds. Each point on the line represents the sensitivity and specificity, calculated for one single threshold.

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall} = \frac{PPV \times Sensitivity}{PPV + Sensitivity} \tag{2.6}$$

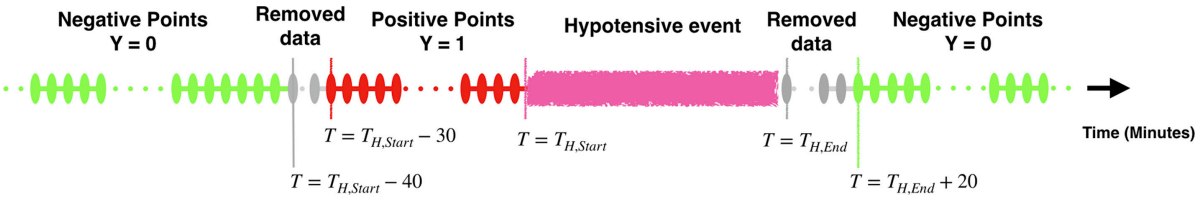


Figure 2.1: Illustration of forward sliding window labelling method. Negative points are predictions that are followed by non-hypotension, thus TN or FP. Positive predictions are predictions that are followed by hypotension, thus TP or FN. The image is reprinted from Moghadam *et al.* [37], with permission from Elsevier.

2.3.2. Forward tumbling window validation

The forward tumbling window (FTW) validation protocol was used by Wijnberge *et al.* [60] to validate HPI scores based on continuous non-invasive blood pressure data during surgery. Nonetheless, this validation protocol can also be applied for validation of HPI on intensive care patients. Data preprocessing consisted of removal of suspected hemodynamic intervention, as described at Section 2.2.1. No imputation of missing values was applied.

Hypotension was defined as a MAP < 65 mmHg for at least one minute. In contrast to other validation protocols, non-hypotension was defined as a MAP > 65 mmHg instead of > 75 mmHg. An alarm was defined as an HPI value > 85 for at least one minute.

Data was labelled with a tumbling window approach. So, every 20-minute window was assigned one label. The time window was 'tumbled' or 'flipped' ahead in time, so windows did not overlap. Data was sequentially labelled from the start to the end of the data timeline of a patient. As long as no alarm was encountered, each past 20 minute window was labelled based on the occurrence of hypotension in that window (as TN or FP). Upon an alarm, a new 20 minute window - starting from the alarm - was forced. Again, its label depended on the occurrence of hypotension in this window (TP or FN). Every next window only started once hypotension had resolved. Performance metrics consisted of sensitivity, specificity, PPV and NPV for an HPI threshold of 85. This threshold of 85 was chosen as it is the threshold of the FDA-approved and commercially available HPI algorithm.

2.3.3. Backward validation

The backward (BW) protocol by Hatib *et al.* [19] was the original method used in the first publication on HPI performance. To date, this is the only validation study on HPI performance in the general ICU population, albeit internal validation. So, only to this protocol could the external PHYSIC cohort be directly compared, as both cohorts include the same patient population.

Data preprocessing consisted of data removal of segments that were suspected to be artefacts (e.g. change of pressure transducer height) , external events or acute events (e.g. sudden blood loss) that were outside the scope of the predictive algorithm. A rate of decrease in MAP of > 0.5 mmHg/s was deemed non-physiological. Every data point that showed a decreasing trend of > 0.5 mmHg/s was excluded from analysis.

A hypotensive event was defined as a section in which each data point showed a MAP < 65 mmHg for a minimal duration of one minute. A non-hypotensive episode was defined as a period which satisfies two conditions: 1) all data points have a MAP of > 75 mmHg for a minimal duration of 30 minutes, and 2) data points are separated from any hypotensive event

by at least 20 minutes.

Labelling of events was performed by the 'backward' methodology. Again, a label (TP, FP, TN or FN) consists of a pair of binary values on prediction and a pair of binary values on outcome. In backward labelling, the actual outcomes are annotated first, i.e. the onset of hypotension. Then, this outcome is paired with a prediction value from a specified duration *'t'* prior to the outcome value. Event prediction samples were defined as the HPI value recorded exactly *'t'* minutes prior to the onset of the hypotensive event. Results are presented separately for each lead time *'t'*, with *'t'* being 5, 10 and 15 minutes. Non-event samples were defined as the mid-point of each 30-minute non-hypotension episode. This would reportedly reduce intraclass correlation. Each individual prediction per 20 seconds was considered to come from nearly the same hemodynamic state. Repetitive information from including all information would have introduced bias, according to Hatib *et al.* [19].

Performance metrics were separately calculated for each lead time *'t'* (5, 10 or 15 minutes). Sensitivity, specificity, positive predictive value, and negative predictive value were calculated for each lead time using a statistically optimal threshold. The optimal threshold was defined to result in the minimal difference between sensitivity and specificity. Other metrics included the ROC curve, with displayed area under the ROC curve (AUROC) and a calibration curve. A calibration curve illustrates the mean rate of event occurrence per HPI range.

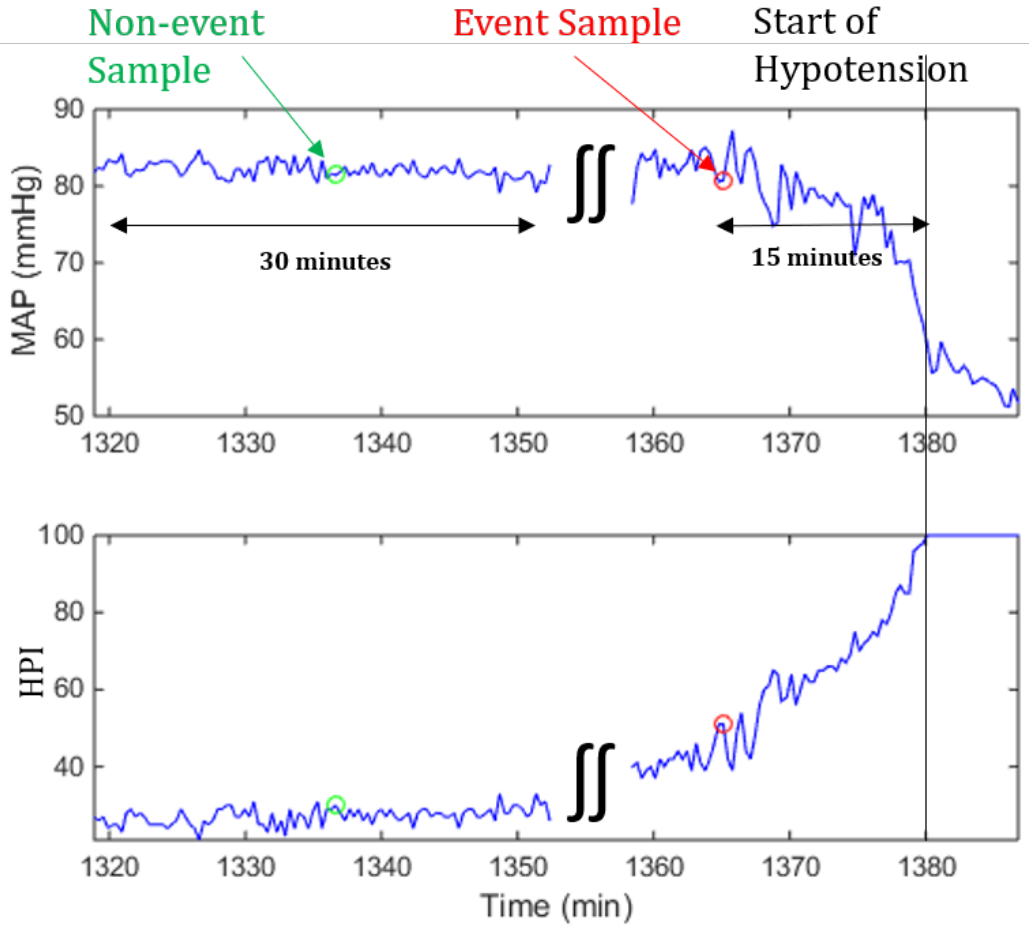


Figure 2.2: Illustration of backward labelling method. The example uses a *'t'* of 15 minutes to pair the onset of hypotension with a prediction. The image was reprinted from Hatib *et al.* [19], with permission from Elsevier.

2.4. Validation types: internal and external validation

An important aspect that contributes to the added value of this thesis is that the HPI model is *externally* validated on the ICU population. In contrast, internal validation is often a first step after model development to test the model on reproducibility and is an indication for generalizability. Internal validation uses data that was kept apart from the originally available development data. As internal validation data originates from the same underlying population as the training data, it is sensitive to confounding. Therefore, internal validation only leads to an initial indication of model generalizability, as it only allows conclusions on the directly underlying population the model is trained on.[48] Generalizability can better be assessed via external validation, as this involves an independent data set.[23] Data for external validation can differ in geographic location (e.g., hospital or country), moment in time, and clinical case type or severity.[49] Thus, external validation evaluates a predictive model in a broader context, which shows true model performance that can be expected upon implementation.

Methods

Methods

The primary goal was to validate the HPI by using a clinically relevant validation protocol, and to compare these results with previously published methods. Therefore, three different validation protocols were applied to the same database of patients admitted on the intensive care unit (ICU). As described below, the primary validation protocol is a newly tailored version of the protocol by Moghadam *et al.* [37] (Section 2.3.1). The other validation protocols are the protocols published by Wijnberge *et al.* [60] (Section 2.3.2) and Hatib *et al.* [19] (Section 2.3.3).

3.1. PHYSIC database

The PHYSIC database was used for validation of the HPI. This database, from AmsterdamUMC – Location Academic Medical Center, consists of 499 adult patients admitted to the intensive care unit (ICU).The PHYSIC was acquired in a different study. Patient inclusion criteria were: age above 18 years, expected minimum stay on ICU longer than eight consecutive hours and that arterial blood pressure was already recorded as part of standard care. Exclusion criteria consisted of: inability to measure blood pressure continuously, a target MAP below 65 mmHg and logistic difficulties around patient transport.

For each included patient, continuous arterial blood pressure data was collected for an intended duration of seven to eight hours. The goal was to measure blood pressure with a five French cannula in the radial artery. If the radial artery was not accessible, the brachial or femoral artery could be used to measure blood pressure. Baseline patient characteristics that were obtained included: age, weight, height, sex, intoxications, medical history, reason for ICU admission and sequential organ failure assessment (SOFA) score. Continuous blood pressure data was acquired using the FloTrac EV1000 hemodynamic monitor (Edwards Lifesciences LTD, Irvine, CA, USA). Hemodynamic variables that were automatically derived from the blood pressure data included: mean arterial blood pressure (MAP), systolic arterial blood pressure, diastolic arterial blood pressure, heart rate and variables derived from the arterial waveform such as cardiac output, cardiac index, stroke volume, SV index, stroke volume variation and pulse pressure variation. All hemodynamic variables were averaged per 20 second window. After data collection, Edwards Lifesciences retrospectively applied their proprietary HPI algorithm on the continuous blood pressure data. HPI values were returned per 20 second window by design and added to the PHYSIC database for the purpose of validation.

The study that led to the formation of the PHYSIC database was conducted with approval of the Medical Ethical Committee of Amsterdam UMC - Location Academic Medical Center under source ID: W18_142#18.176. The study was also included in the Netherlands Trial Register under ‘NTR7349’. Consent was obtained for every patient included in the database. Data was anonymized with its key only available to main researchers.

3.2. Other materials

Validation protocols of Moghadam *et al.* [37] and Hatib *et al.* [19] were reproduced in the Python programming language version 3.9.4 (The Python Software Foundation, Wilmington, DE, USA).[54] This included data preprocessing, labelling, calculation performance metrics and creation of figures. Frequently used libraries were Pandas[36] version 1.2.5, Numpy[17] version 1.21.0, Scikit-learn[40] version 0.24.2 and Matplotlib[22] version 3.4.2. A full list of required libraries is listed in the 'Requirements' text file with the scripts. For the validation protocol of Wijnberge *et al.* [60] the original code could be reused.Therefore, this analysis performed in MATLAB (The MathWorks Inc., Natick, MA, USA)[32], version 9.5.0 (R2018b). For calculation of extra performance metrics, additional code was added to the original scripts.

A secure digital workspace on a remote server was used for coding and testing with patient data (Azure DRE, anDREa, Nijmegen, The Netherlands). Offline version management was performed via Git[6] version 2.21.0.windows.1. Python scripts for performed validation protocols are available online (via https://github.com/mpligtenberg/HPI-validation). Other code and data is available upon request.

3.3. Primary validation protocol: Forward sliding window

A newly tailored version of the forward sliding window (FSW) was used as the primary validation protocol in this thesis. This protocol approximates real-time clinical application of the HPI by using single predictions in the calculation of performance metrics. Moghadam *et al.* [37] executed the validation protocol on a different algorithm with different constraints (Section 2.3.1). Therefore, the validation protocol was tailored to the application of the HPI on the PHYSIC database, as detailed below.

3.3.1. Data preprocessing

Missing data was interpolated linearly if the missing interval was less than five minutes. Data segments separated by missing intervals of 5 minutes or larger treated as separate segments in analysis.

Data points with bad signal quality or showing non-physiological values or changes in values were annotated as bad data points. Labels based on data segments containing bad data points were excluded from analysis.

Bad signal quality was annotated by a proprietary detection algorithm by Edwards Lifesciences. Non-physiological blood pressure values were annotated, according to the physiological ranges that were used by Moghadam *et al.* [37] (Section 2.3.1). Non-physiological changes in MAP measurements were annotated by using the same rules as used in previous validation by Hatib *et al.* [19] and Wijnberge *et al.* [60]. A decrease of MAP greater than 0.5 mmHg/s was regarded as non-physiological as this would result from an external factor, rather than the changing state of the patient.[19] A MAP increase of 5 mmHg in 20 seconds from any baseline or 8 mmHg in 2 minutes form a baseline of less than 70 mmHg was assumed to result from a hemodynamic intervention, thus non-physiological.[19,60] Removal of spikes based on the variance, as in the original protocol of Moghadam *et al.* [37], was not feasible.

3.3.2. Alarm definition

An alarm is defined by a combination of an HPI value and the alarm threshold. An alarm is an HPI value that exceeds the alarm threshold. No minimal duration for HPI value above the threshold is applied. Multiple HPI threshold values are used to calculate performance metrics. This allows interpretation by the individual reader on optimum alarm thresholds. The optimal HPI threshold depends on patient characteristics and individual user preferences when to be alarmed or when to initiate pro-active treatment. Therefore, performance was presented for thresholds that include all multiples of five and additional statistically optimal thresholds: maximum f1-score, sum of sensitivity and specificity (Youden Index, as used by Wijnberge *et al.* [60]) and minimal difference between sensitivity between sensitivity and specificity (as used by Hatib *et al.* [19]).

3.3.3. Hypotension definition

A hypotensive event was defined as a MAP < 65 mmHg for a period of at least one minute. Non-hypotension was defined as MAP ≥ 70 mmHg.

3.3.4. Prediction window

Every HPI value is regarded to cast a prediction over the succeeding 15 minutes, as the HPI model was developed with waveforms up until 15 minutes prior to onset of hypotension. Therefore, a sliding window with a duration of 15 minutes was used to assess HPI performance. The correctness of every alarm is based on the occurrence of hypotension in this window. A leading neutral buffer of 5 minutes was used, so that alarms with a time-to-hypotension of 15 to 20 minutes were not labelled as a FP. A washout period of 30 minutes was used. A washout period reduces the influence of repetitive hypotension onsets that inflate the number of hypotension predictions, whereas they could be regarded as one episode.

3.3.5. Performance metrics

Metrics for discriminative performance (sensitivity, specificity, PPV, NPV and F1-score) were calculated for each alarm definition discussed above at Section 3.3.2. Discriminative performance was further evaluated using the Receiver Operating Characteristic (ROC) curve (i.e., a graph of specificity against sensitivity for all alarm thresholds) and precision-recall (PR) curve (i.e., a graph of PPV against sensitivity for all alarm thresholds). Each point on the line represents the sensitivity and specificity or PPV, calculated for one single threshold. As an addition to conventional curves, the colour of the line indicates the alarm threshold used to calculate the values. Area under the ROC and PR curves (AUROC and AUCPR, respectively) were also used as a summarizing metric.

Performance on calibration was illustrated using a calibration curve and summarized by the Matthews Correlation Coefficient[35].

In addition to discrimination and calibration, timeliness of the alarm is fundamental for the added clinical value of the HPI mode. Timeliness of the algorithm was graphically illustrated by showing the sensitivity for each time interval *'t'* between prediction and onset of hypotension, i.e. the time-to-hypotension *'t'*. By evaluating the sensitivity for all predictions made *'t'* minutes

before the onset of hypotension, effectively, backward validation is applied. Sensitivity was plotted against time-to-hypotension, as early as 30 minutes before onset of hypotension.

Also, timeliness of HPI was assessed via distribution of time-to-hypotension of alarms. This is the duration between alarm and onset of hypotension. The distribution was presented in a histogram. This was done by two different methods, resulting in two histograms. The backward method evaluated the time interval between onset of hypotension and the last of consecutive alarms prior to hypotension. The forward method evaluated the time interval between the earliest alarm within the prediction window and the onset of hypotension. If a washout period was in effect within the prediction window, this hypotensive event was not included in the forward distribution of time-to-hypotension durations. Median values with interquartile range (IQR) were also reported as summarizing metric.

3.4. Other protocols

Previously published protocols were also applied to the PHYSIC database to enable comparison with the FSW protocol. Backward (BW) validation was performed according to the protocol of Hatib *et al.* [19]. Forward tumbling window (FTW) validation was performed according to the protocol by Wijnberge *et al.* [60]. The BW and FSW protocols were performed as described in Chapter 2. However, the same performance metrics as in the primary protocol were used, in order to directly compare the labelling methods.

3.5. Subgroup analyses

Clinically relevant subgroups of patients were evaluated individually on HPI performance. Subgroup performance was compared to non-subgroup performance. Subgroups that were analysed were: patients with cardiogenic shock, distributive shock, admission post cardiothoracic surgery and admission due to a subarachnoid hemorrhage.

3.6. Exploratory analyses

To evaluate bias induced by different protocol elements, different variations of primary protocol were applied. Each variation was applied individually to enable clear comparison with the baseline protocol. Different elements consisted of: a non-hypotension threshold of a MAP greater than 65, 70 and 75 mmHg; a predicted window of an alarm of 5, 10, 15 and 20 minutes; a leading neutral buffer of 0, 10 and 20 minutes; a washout period of 0, 10, 20 and 30 minutes; downsampling of non-hypotensive events by 20% to show the result of class imbalance; a scatter plot of HPI against MAP for single predictions.

Results

Results

4.1. Data preprocessing

Data preprocessing of the primary FSW protocol resulted in 212 splits on missing data intervals of greater than five minutes. A total of 7345 imputations were performed on shorter missing data intervals to a total of 920987 data points. The amount of excluded data points per label is shown in Appendix B.

4.2. Baseline patient characteristics

Baseline characteristics of patients included in the PHYSIC database are presented in Table 4.1. The median monitoring duration was 7 hours and 21 minutes.

Table 4.1: Baseline patient characteristics in the PHYSIC database.

Baseline parameters	
Total number of patients	499
Sex, male, n (%)	327 (66)
Age, years, mean (sd)	61 (14)
Number of patients older than 65 years, n (%)	221 (44)
Weight (kg), mean (sd)	82.97 (19.5)
Height (cm), mean (sd)	174 (9.9)
BMI, mean (sd)	27 (6)
SOFA score, mean (sd)	10 (3)
Vasoactive medication during measurements, n (%)	302 (61)
Mechanical ventilation, n (%)	358 (72)
Measurement details	
Monitoring time per patient (minutes), median [Q1-Q3]	441 [411 – 962]
Number of daytime measurements, n (%)	305 (61)
Number of night-time measurements, n (%)	194 (39)
Reason of ICU admission	
Respiratory failure, n (%)	57 (11)
Neurological disease, n (%)	82 (16)
Subarachnoid haemorrhage , n (%)	51 (10)
Sepsis, n (%)	38 (8)
Cardiac shock/other cardiac, n (%)	19 (4)
Postoperative after surgery, n (%)	216 (43)

Table 4.1 continued from previous page

Baseline parameters	
Cardiothoracic surgery, n (%)	199 (40)
Assigned shock groups	
Cardiogenic shock, n (%)	66 (13)
Distributive shock, n (%)	94 (19)
Hypovolemic shock, n (%)	12 (2)
Obstructive shock, n (%)	2 (0.4)
Combination type of shock, n (%)	32 (6)
Nonshock classification, n (%)	293 (59)

Statistic presented as mean (standard deviation), median [first quartile, third quartile], or number of patients (%). Abbreviations: *MAP*, mean arterial pressure; *BMI*, body mass index; *SOFA*, sequential organ failure assessment.

4.3. Primary analysis: Forward sliding window validation

For the tailored FSW protocol, the AUCPR was 0.59 and the AUROC was 0.87. PR curve, ROC curve and Calibration are presented in Figure 4.1, 4.2, and 4.3. Optimal statistical thresholds for the validation protocol were 94 for maximum f1 score, 64 for maximum Youden Index and 60 for minimal difference between sensitivity and specificity. Performance across all calculated thresholds is described in Appendix C.

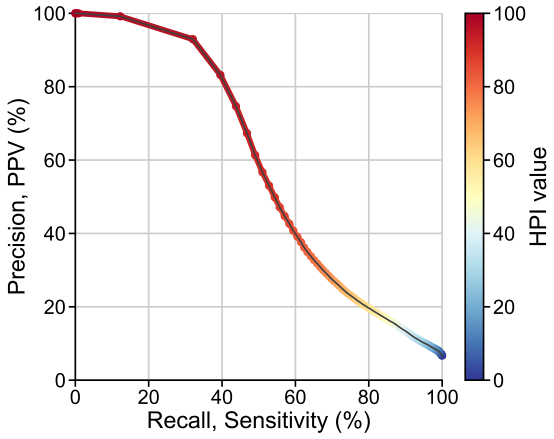


Figure 4.1: PR curve of FSW protocol

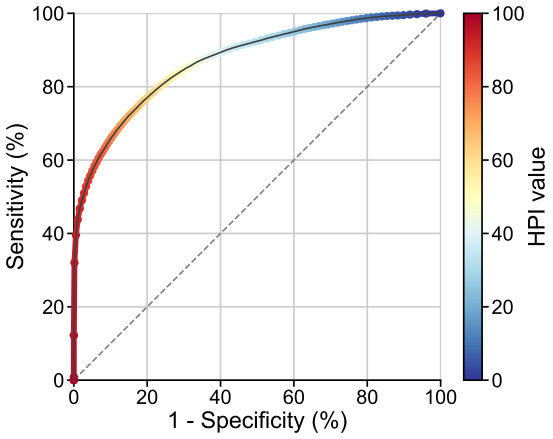


Figure 4.2: ROC curve of FSW protocol

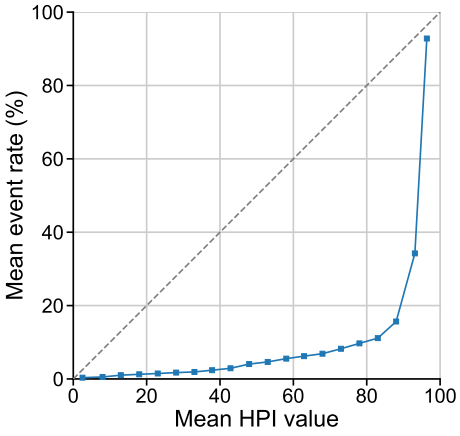


Figure 4.3: Calibration curve for the FSW protocol, with an alarm threshold of 85

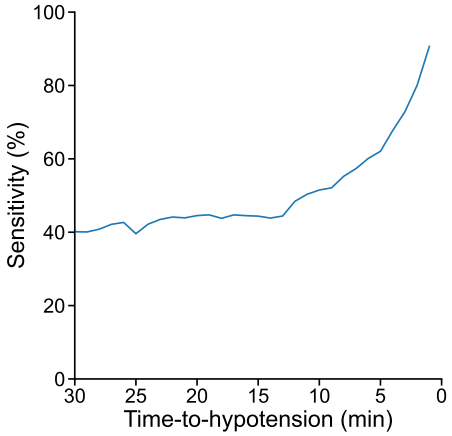


Figure 4.4: Sensitivity against time-to-hypotension, using an alarm threshold of 85.

The median time-to-hypotension [IQR] was 3.3 [1.0 to 7.3] minutes for backward analysis (Figure 4.5) and 14.3 [6.5 to 15] minutes for forward assessment (Figure 4.6), by using an HPI threshold of 85. No alarm went off for 35 of the 759 hypotensive events of which the full prediction window was available. The alarm went off in the entire prediction window in 130 of 982 evaluated hypotensive events. The HPI became more sensitive as the predictions are made closer to the onset of hypotension, as illustrated in Figure 4.4. From 30 to 13 minutes before onset of hypotension, the sensitivity increased from 40% to 45% Thereafter, sensitivity increases more rapidly. An increase in HPI threshold from 65 to 95 resulted in a decrease of time-to-hypotension for all quartile values (Table 4.2). All distributions are included in Appendix H.

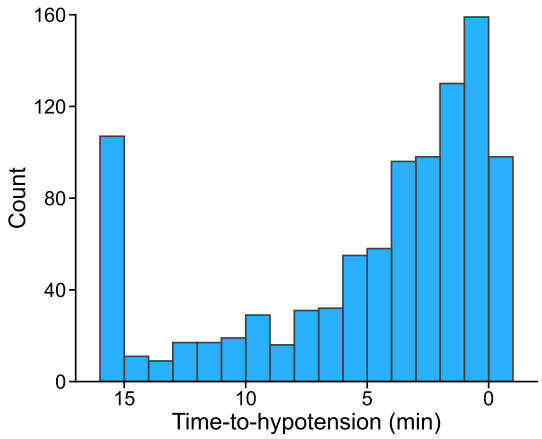


Figure 4.5: Backward timeliness assessment. Distribution of time to hypotension in backward timeliness assessment, using an alarm threshold of 85. For backward assessment, the alarm is defined as the lasts of consecutive alarm prior to hypotension. Bars include the higher edge value end exclude the lower edge value. The bar for a time-to-hypotension higher than 15 min indicates the amount of hypotensive events that were preceded by 15 minutes of non-stop HPI alarms. The bar with a negative time value indicates the number of hypotensive events without an alarm in the 20 second before onset. The total count is 982.

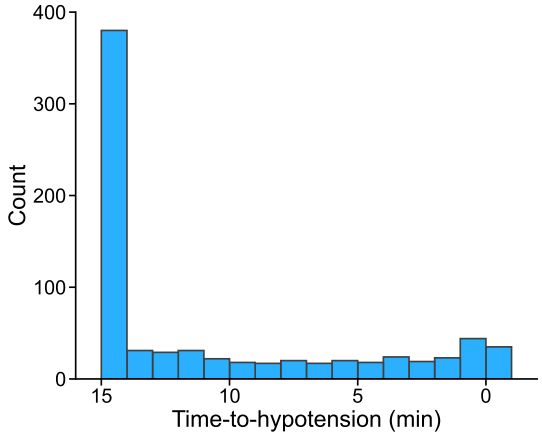


Figure 4.6: Forward timeliness assessment. Distribution of time to hypotension in forward timeliness assessment, using an alarm threshold of 85. For forward assessment, the alarm is defined as the first HPI value above threshold in the prediction window. Bars include the higher edge value end exclude the lower edge value. The bar with a negative time-to-hypotension indicates the number of hypotensive events that were not preceded by any alarm. The total count is 759.

Table 4.2: Time-to-hypotension across thresholds for forward and backward timeliness assessment. Time is stated in minutes.

	Forward				**Backward**			
HPI threshold	65	75	85	95	65	75	85	95
25th percentile	12	9.6	6.5	1.7	2.3	1.6	1.0	0
50th percentile	15	15	14.3	8.3	6	4.3	3.3	1.0
75th percentile	15	15	15	14.3	13	11	7.3	2.7

4.4. Forward tumbling window validation

For the FTW protocol, the AUCPR was 0.95 and the AUROC was 0.97. The PR curve and ROC curve are presented in Figure 4.7 and 4.8. The optimal statistical threshold for the validation protocol was 95 for all statistical optimums: maximum f1-score maximum Youden Index, and minimal difference between sensitivity and specificity. The median time-to-hypotension [IQR] was 2.7 [1 to 6.3] min. Performance across all calculated thresholds is described in Appendix D.

4.5. Backward validation

For the BW protocol with lead times of 5, 10 and 15 minutes, the AUCPR was 0.99, 0.99 and 0.97, respectively. The AUROC was 0.99, 0.98 and 0.95, respectively. PR curve, ROC curve and Calibration are presented in Figure 4.9, 4.11, and 4.13. For more detail, zoomed figures were created (Figure 4.12, 4.10). For lead times of 5, 10 and 15 minutes, the optimal statistical thresholds for the validation protocol with maximum f1-score were 52, 57 and 70, with maximum Youden Index were 64, 79 and 65 and with minimal difference between sensitivity and specificity were 51, 71 and 61. Performance across all calculated thresholds is described in Appendix E.

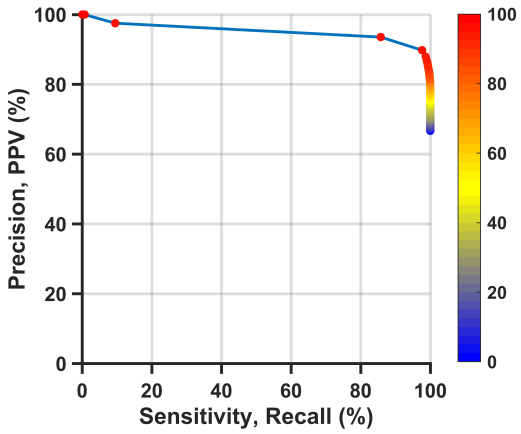


Figure 4.7: PR curve of FTW protocol

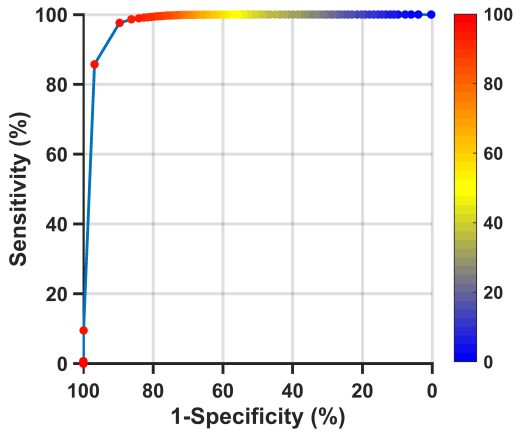


Figure 4.8: ROC curve of FTW protocol

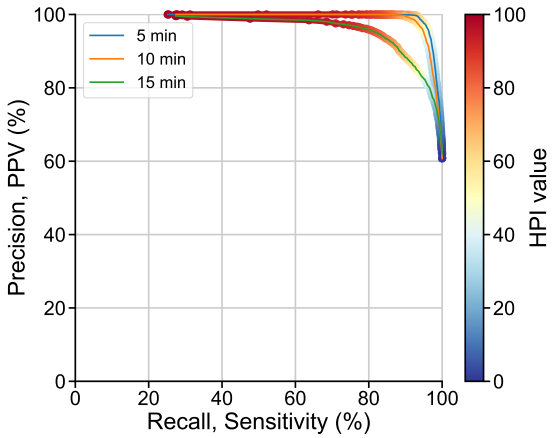


Figure 4.9: PR curve of BW protocol, for different lead times *'t'*

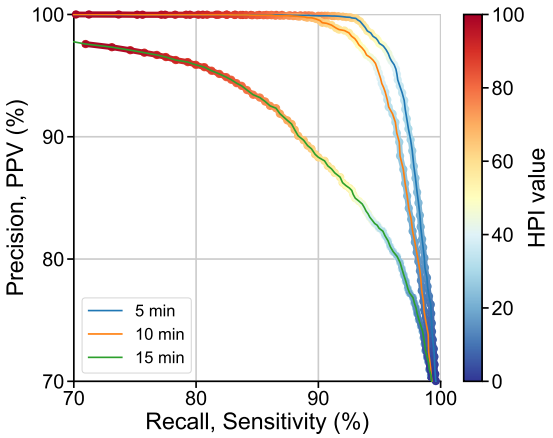


Figure 4.10: Zoomed PR curve of BW protocol, for different lead times *'t'*

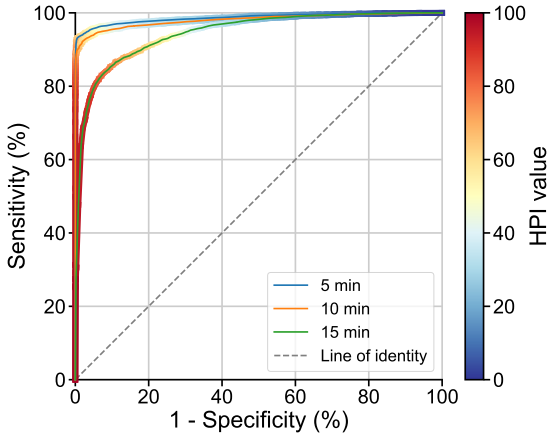


Figure 4.11: ROC curve of BW protocol, for different lead times *'t'*

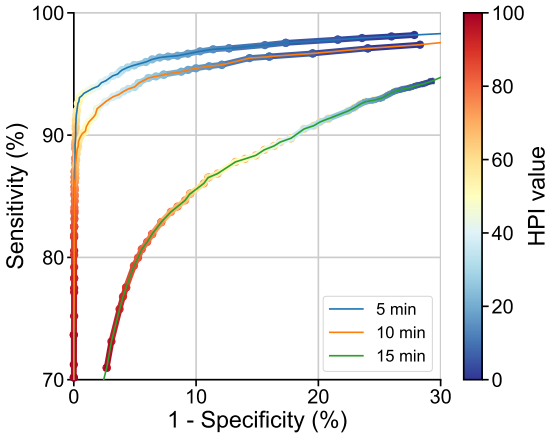


Figure 4.12: Zoomed ROC curve of BW protocol, for different lead times *'t'*

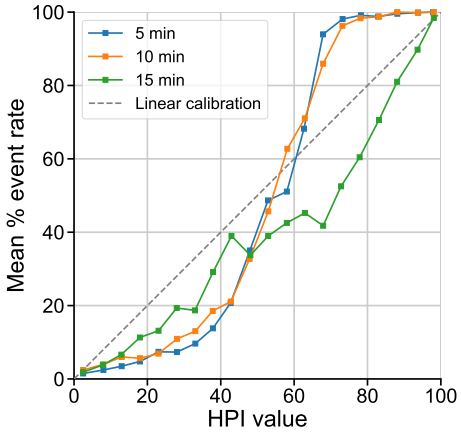


Figure 4.13: Calibration curve of BW protocol, for different lead times *'t'*

4.6. Overview of all protocols, for alarm threshold 85

Table 4.3: Overview of performance metrics for all validation protocols, using an alarm threshold of 85. For the BW protocol, lead time *'t'* is stated in round brackets.

	FSW	FTW	BW (5 min)	BW (10 min)	BW (15 min)
Se	0.59	1.00	0.86	0.83	0.81
Sp	0.94	0.77	0.99	0.99	0.94
PPV	0.41	0.83	1.0	1.0	0.96
NPV	0.97	1.00	0.82	0.79	0.76
AUROC*	0.873	0.972	0.987	0.981	0.946
AUCPR*	0.585	0.951	0.993	0.990	0.965
MCC	0.448	0.794	0.837	0.811	0.735
no. of AHE*	982	7654	6072	6025	5972
TP	21346	7628	5207	5010	4818
FP	31050	1566	2	2	217
TN	470562	5107	3901	3901	3686
FN	14573	22	865	1015	1154

Fields with a (*) do not depend on the threshold value. Abbreviations: *Se*, sensitivity; *Sp*, specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *AUROC*, area under the ROC curve; *AUCPR*, area under the PR curve; *MCC*, Matthews correlation coefficient.

4.7. Secondary analyses

Secondary analyses were performed using the primary validation protocol or single parameter variations on the primary protocol. The primary validation protocol applied sliding forward validation with a prediction window of 15 minutes, leading neutral buffer of 5 minutes, washout period of 30 minutes and a non-hypotension threshold of > 70 mmHg.

All ROC curves and Calibration curves of the secondary analyses are placed in Appendix F.

4.7.1. Subgroup analyses

Subgroup performance with an alarm threshold of 85 is presented in Table 4.4. PR curves are presented for each subgroup analysis (Figure 4.14, 4.15, 4.16, 4.17).

For patients admitted after cardiothoracic (CAPU) surgery versus other patients, the AUCPR was increased (0.61 vs 0.56), but the AUROC decreased (0.85 vs 0.89). For patients with a subarachnoid haemorrhage (SAH) versus non-SAH patients, the AUCPR was decreased (0.28 to 0.59).

The PPV value at 100% sensitivity is the percentage of hypotension labels in the data subset. Therefore this value represents the class balance between hypotension and non-hypotension.

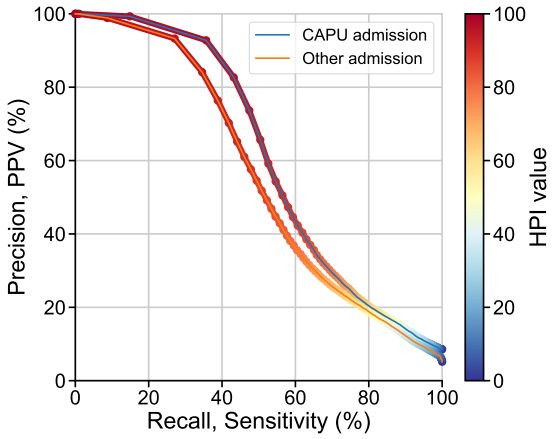


Figure 4.14: PR curves for patients admitted after cardiothoracic (CAPU) surgery

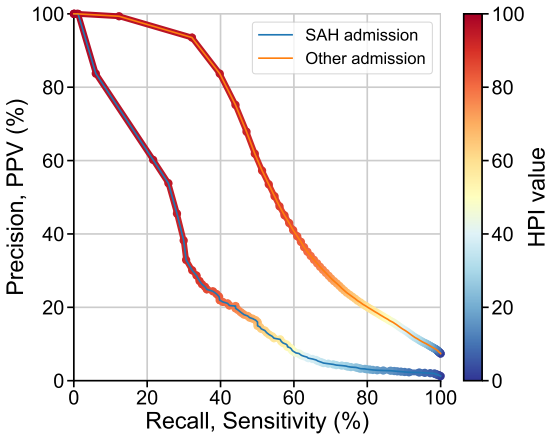


Figure 4.15: PR curves for patient subgroups with and without a subarachnoid haemorrhage (SAH)

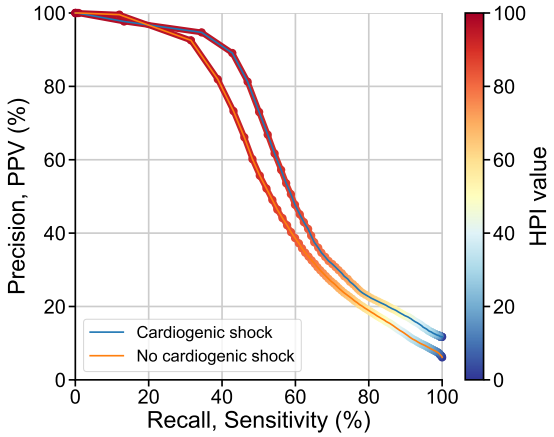


Figure 4.16: PR curves for patient subgroups with and without a cardiogenic shock

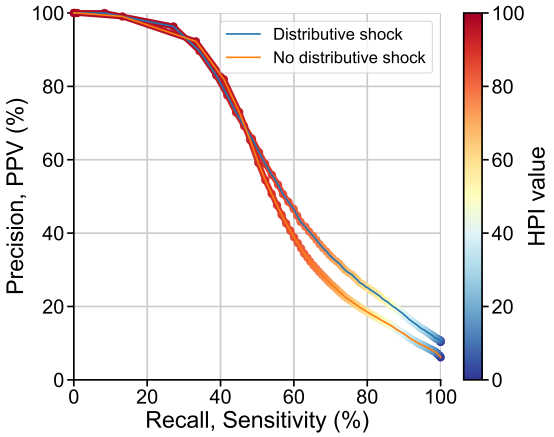


Figure 4.17: PR curves for patients with and without a distributive shock

Table 4.4: Performance per subgroup, with an alarm threshold of 85.

Subgroup	**CAPU Admission**		**SAH admission**		**Cardiogenic shock**		**Distributive shock**	
	Yes	No	Yes	No	Yes	No	Yes	No
Se	0.63	0.55	0.38	0.60	0.62	0.59	0.57	0.605
Sp	0.90	0.96	0.99	0.93	0.89	0.94	0.94	0.94
PPV	0.39	0.45	0.24	0.41	0.43	0.40	0.51	0.39
NPV	0.96	0.98	0.99	0.97	0.95	0.97	0.95	0.97
AUROC	0.848	0.892	0.883	0.864	0.829	0.876	0.853	0.875
AUCPR	0.608	0.558	0.280	0.591	0.631	0.576	0.614	0.583

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *AUROC*, area under the ROC curve; *AUCPR*, area under the Precision Recall curve; *CAPU*, cardiothoracic surgery; *SAH*, subarachnoid haemorrhage.

4.7.2. Non-hypotension definition

Decreasing the minimal MAP in non-hypotension definition from 75 to 65 mmHg reduces specificity, PPV, AUCPR and AUROC (Figure 4.18, Table 4.5). In particular, the AUCPR decreased from 0.82 to 0.31, respectively.

Table 4.5: Performance metrics for different thresholds of non-hypotension, with an alarm threshold of 85.

	MAP		
	> 65 mmHg	> 70 mmHg	> 75 mmHg
Se	0.59	0.59	0.59
Sp	0.85	0.94	0.999
PPV	0.21	0.41	0.98
NPV	0.97	0.97	0.96
AUROC	0.822	0.873	0.930
AUCPR	0.307	0.585	0.816

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *AUROC*, area under the ROC curve; *AUCPR*, area under the Precision Recall curve.

4.7.3. Prediction window duration

For prediction window durations of 5 and 20 minutes, the AUCPR was 0.63 and 0.54, respectively. An increase in the prediction window resulted in an increase in PPV, but a decrease of sensitivity (Figure 4.19, Table 4.6).

Table 4.6: Performance metrics for different prediction windows, supplemented with a leading neutral buffer to a total of 20 minutes, for an HPI threshold of 85.

	Prediction window			
	5 min	10 min	15 min	20 min
Se	0.75	0.65	0.59	0.56
Sp	0.94	0.94	0.94	0.94
PPV	0.24	0.35	0.41	0.45
NPV	0.99	0.98	0.97	0.96
AUROC	0.930	0.894	0.873	0.860
AUCPR	0.632	0.600	0.585	0.584

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *AUROC*, area under the ROC curve; *AUCPR*, area under the Precision Recall curve.

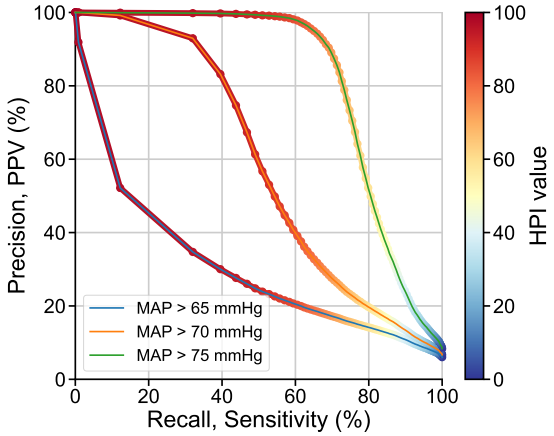


Figure 4.18: PR curves for different non-hypotension definitions

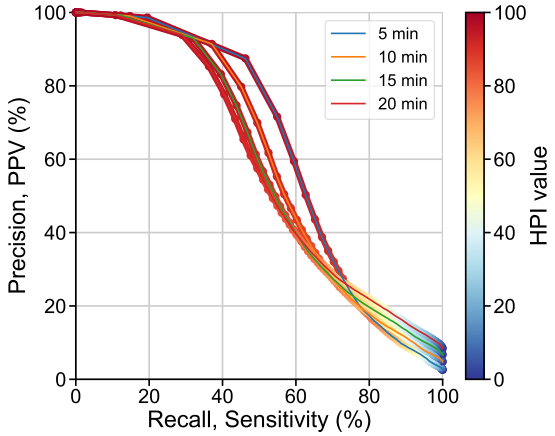


Figure 4.19: PR curves for different prediction window durations

4.7.4. Leading neutral buffer zone

The effect of a leading neutral buffer zone between negative and positively labelled points prior to hypotension relatively affected the PPV the most. For a neutral buffer of 0 min and 20 min the PPV was 0.40 and 0.43, respectively (Figure 4.21, Table 4.7).

Table 4.7: Performance metrics for different leading neutral buffer periods, for an HPI threshold of 85.

	Neutral buffer		
	0 min	10 min	20 min
Se	0.59	0.60	0.60
Sp	0.94	0.94	0.94
PPV	0.40	0.42	0.43
NPV	0.97	0.97	0.97
AUROC	0.871	0.875	0.878
AUCPR	0.580	0.590	0.600

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *AUROC*, area under the ROC curve; *AUCPR*, area under the Precision Recall curve.

4.7.5. Washout period

Reducing the washout period from 30 min to 0 min resulted in an increased PPV from 0.41 to 0.64, using an alarm threshold of 85 (Figure 4.21, Table 4.8). An labelling example including a washout period is illustrated in **??**.

Table 4.8: Performance metrics for different washout periods, for an HPI threshold of 85.

	Washout period			
	0 min	10 min	20 min	30 min
Se	0.72	0.63	0.61	0.59
Sp	0.93	0.93	0.94	0.94
PPV	0.64	0.49	0.44	0.41
NPV	0.95	0.96	0.97	0.97
AUROC	0.9089	0.883	0.876	0.873
AUCPR	0.788	0.659	0.611	0.585

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *AUROC*, area under the ROC curve; *AUCPR*, area under the Precision Recall curve.

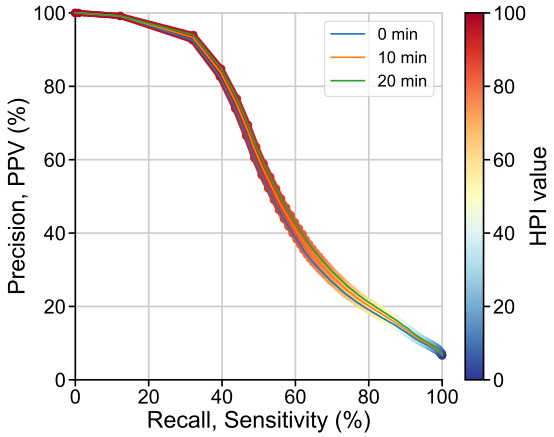


Figure 4.20: PR curves for different leading neutral buffer durations

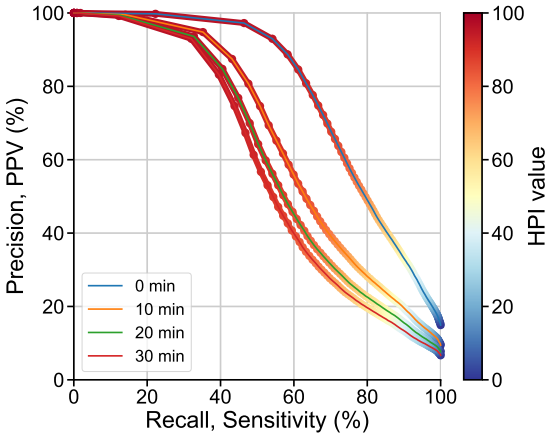
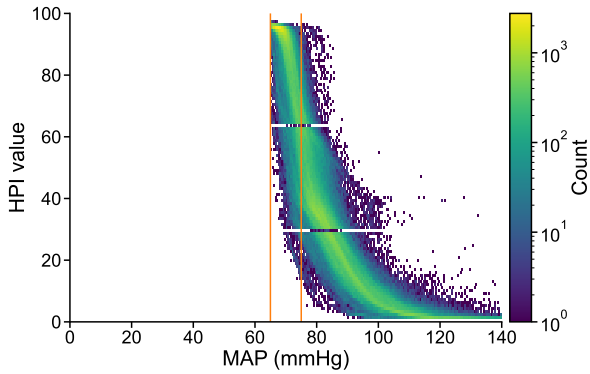


Figure 4.21: PR curves for different washout period durations

4.7.6. HPI vs MAP

HPI is plotted against MAP for single predictions followed by non-hypotension (Figure 4.22) and for single predictions followed by hypotension (Figure 4.23).



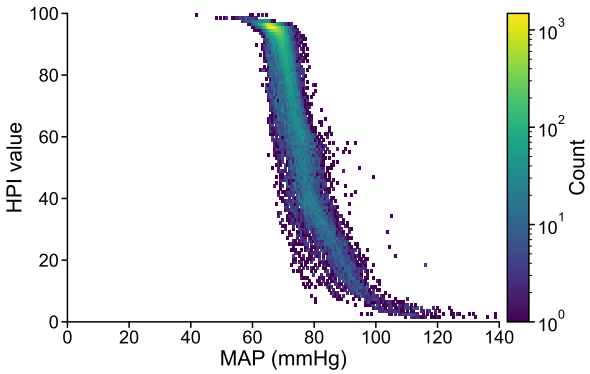


Figure 4.22: HPI vs MAP: Negative points (TN and FP) **Figure 4.23:** HPI vs MAP: positive points (TP and FN)

Discussion

Discussion

This thesis shows that validation methodology is of utmost importance in validation of the Hypotension Prediction Index (HPI). The validation protocol should mimic the real-life implementation of the predictive algorithm in labelling of the data points. Also, the protocol should integrate an appropriate washout period, clinically relevant blood pressure values for hypotension and non-hypotension, and a true representations of hypotension and non-hypotension occurrence.

5.1. Primary analysis

In this first external validation of the HPI model on general ICU population, the HPI model showed good predictive performance in predicting hemodynamic instability. The results demonstrate generalizability of the HPI model to the ICU population, as the ROC curves showed decent overall discriminative performance for every protocol. However, the use of the primary forward sliding window (FSW) protocol resulted in reduced performance metrics when compared to previously published labelling methods for the HPI model. Especially the positive predictive value (PPV) for the FSW protocol was significantly reduced when compared to the forward tumbling window (FTW) protocol and backward (BW) protocol.

A major cause of the reduced PPV and calibration in the FSW protocol could be the lower ratio of hypotension labels versus non-hypotension labels. The PPV depends on this ratio: more non-hypotension leads to increased probability on FPs, which reduces the PPV. The FTW and BW protocol increased the hypotension ratio, by either undersampling the non-hypotension data or by grouping the points into time windows. For example, one label (TN or FP) was assigned per 20 or 30 minutes of non-hypotension data points. In contrast, the primary FSW protocol labelled each single prediction per 20 seconds as TN or FP. Thereby, the FSW protocol labels more non-hypotension, which increases the difference between numbers of hypotension and non-hypotension labels. As this so-called ‘class imbalance’ leant more towards non-hypotensive events, the PPV was reduced as a result (Table 2.1). I believe the reduced PPV value of the FSW protocol depicts a more realistic value for HPI predictions, as the FSW protocol mimics the interpretation bed-side HPI values. In calculation of PPV, a more realistic ratio between occurrence of hypotension and non-hypotension is used.

The HPI showed poor calibration in the primary FSW validation protocol. The calibration curve does not show a proportional increase in rate of hypotension over the full range of HPI values (Figure 4.3). So, the amount of increase in HPI value itself was not proportional to the increase in hypotension occurrence. This means that the interpretation of the continuous HPI value may not be intuitive. However, a sharp inflection point in the rate of hypotension around an HPI value of 85 was observed: HPI values below 85 showed a low rate of hypotension and values above 85 showed a high rate of hypotension. This could coincide with a good discriminative

performance when an alarm threshold of 85 is used. Calibration results of the FSW protocol are in contrast with calibration resulting from the BW protocol (Figure 4.13) and the article of Hatib *et al.* [19] as they report a sigmoid calibration curve.

In evaluation of the timeliness of the HPI alarms, sensitivity increases as predictions are made closer to the onset of hypotension, especially from 13 minutes to onset of hypotension. The HPI also alarmed more time in advance when lower alarm thresholds were used. In the distribution of forward timeliness, the initial peak in 'alarm counts' between 14 and 15 minutes to hypotension (Figure 4.6) corresponds with the sensitivity of the HPI at 15 minutes (Figure 4.4). In addition, the differences in distributions of time-to-hypotension between forward and backward assessment seem large. This indicates that the first alarm and the onset of hypotension were often separated by a non-alarm.

5.2. Secondary analyses

The HPI model showed better performance in particular subgroups of patients. For patients with a cardiogenic shock and patients admitted after cardiothoracic surgery, the HPI model showed a better trade-off between sensitivity and PPV. HPI performed poorly on patients admitted due to a subarachnoid haemorrhage (SAH). Poor performance may be caused by the low prevalence of hypotension defined as a MAP < 65 mmHg in patients admitted with SAH, versus patients without SAH (1% vs 8% of labels were predictions on hypotension). This could be explained by the higher target MAP in patients admitted due to a SAH, compared to other patients.[42,44]

For validation on the general ICU population, the definition of non-hypotension heavily influences performance results. Lowering the non-hypotension definition from a MAP > 75 mmHg to a MAP > 65 mmHg for labelling of predictions has an extreme effect on the reported PPV, as seen in the PR curves. Many FPs arise in the MAP range of 65 to 75 mmHg in the FSW protocol, as Wijnberge *et al.* [60] also reported. Exclusion of this 'twilight' hypotension zone in analysis, as performed in the BW protocol, polarises the hypotension and non-hypotension points artificially (Figure 4.22). This polarisation lead to easier discrimination between hypotension and non-hypotension (Figure 4.18). Artificial polarisation could be justified during the design of a predictive model. However, for validation, the clinical scenario should be simulated. In clinical practice, the HPI would be used for all MAP values. Therefore, this 'twilight' blood pressure range should also be included in the definition of non-hypotension.

Surprisingly, prediction window size only minimally affected the results on HPI performance in the FSW protocol. The subtle increase in PPV by enlarging the prediction window can be explained as a logical result of increasing the number of positive points. PPV depends on TPs from the positive class and FPs from the negative class. In general, an increase in the ratio between positive and negative points results in an increase in PPV. Differences in performance results do not hint towards any optimal prediction window duration for validation nor towards an expected time to hypotension upon alarm.

A leading neutral buffer duration only minimally affects the results on HPI performance in FSW protocol. This is in contrast with the idea that this buffer would temporally polarize positive and negative points and thus would bias model performance. Possibly, temporal polarisation is limited when using a prediction window of 15 minutes, because most alarms already occur within the prediction window. Results on model performance using smaller prediction windows

might be more influenced by leading neutral buffer size.

Omission of a washout period resulted in an increase in PPV, which may be caused by the increased number of hypotensive event when no washout period is used. Without any washout period, 15% of all labels describe hypotensive events (TP or FN), whereas with a washout period of 30 minutes, this is only 7%. The difference in class balance thus contributes to the change in PPV. In clinical practice, prediction of recurrent hypotensive events could be deemed less valuable, as the patient would be monitored more closely during this recovery period.

5.3. PHYSIC database validity

External validation results on the PHYSIC ICU patient database with the backward (BW) validation protocol are similar to the internal validation results reported by Hatib *et al.* [19]. This indicates that the results presented by Hatib *et al.* [19] are generalizable and that the model is not 'over trained' . The results were similar in terms of ROC curves for predictions made 5, 10, and 15 minutes before onset of hypotension. The calibration curves show similar calibration for predictions made 5 and 10 minutes before onset of hypotension. Despite Hatib *et al.* [19] reporting a sigmoid curve for predictions of 15 minutes before onset of hypotension, external validation with BW protocol indicated under calibration and linear calibration.

It must be noted that only 15% of the internal validation cohort by Hatib *et al.* [19] consisted of ICU patients. The rest of the cohort consisted of surgical patients. Unfortunately, no internal validation results were presented on HPI performance on ICU patients alone.

5.4. Strengths and limitations

Relevantly strong areas in methodology of this thesis were identified. Firstly, this includes the application of multiple and previously reported validation protocols on the same PHYSIC data set. Thereby, differences between the three different protocols could be accurately shown. In addition, by repeating the protocol of Hatib *et al.* [19], the results on both FSW and FTW protocols could be related to previously published results on the HPI. It is clear that the choice of validation protocol is a great determinant of the results on model performance. Therefore, a critical view on reported validation methodology is crucial, prior to accepting the results.

Secondly, performance metrics that display clinically useful information were added to the previously published protocols. The infrequently used PR-curve provides the clinical reader with necessary information for bed-side implementation of the HPI: How sensitive is the alarm? What is the probability of hypotension after an alarm? Additionally, a coloured threshold indicator was added to ROC and PR curves, as a novelty to improve information transfer to the reader. This could aid the estimation of added clinical value of the HPI by the clinical reader.

Lastly, timeliness was assessed via two time-to-hypotension distributions. The variable time is of utmost importance for the potential benefit of an alarm on impending hypotension. Conventionally, timeliness is only presented via median and quartile range values. However, the usefulness of these values were regarded as suboptimal. Therefore, two distributions of time-to-hypotension were used to increase information transfer on timeliness of the HPI model.

The clinical interpretation and generalization of the presented results may be hampered by sev-

eral factors. Firstly, the FSW protocol does not directly mimic the pro-active treatment protocol by using the HPI. It only provides performance metrics on individual predictions. Therefore, direct clinical generalizability of the primary results could be arguably reduced. Clinical application of the HPI would involve only one initiation of pro-active treatment episode of consecutive alarms. Therefore a series of consecutive alarms could also be labelled as a single prediction, as in the FTW protocol. This approach is used byin the FTW protocol. One could argue that labelling individual predictions within the prediction window, as as in the FSW protocol, is irrelevant as a clinician would treat every single alarm. An advantage of this approach is that every no-alarm that is followed by hypotension in the next 15 minutes affects the sensitivity. The FTW protocol only evaluates the first initial alarm and does not take any HPI values into account until the onset of hypotension. So, the validation protocol remains a statistical abstraction, which is subject to design choices. A validation protocol often cannot completely mimic actual clinical use of the predictive model.

Secondly, data preprocessing for the FTW validation protocol was different than for the FSW and BW protocols. Changing the structure of the authors' original code to equal data processing was considered to be outside the time scope of this project. However differences in data preprocessing was regarded as a minimal influence on the results, because of the nature of the windowed labelling approach. By assigning labels to time windows instead of single prediction values, the influence of missing data points is reduced.

Thirdly, in the definition of non-hypotension for the FSW protocol the range of MAP values between 65 mmHg and 70 mmHg was excluded from analysis. As stated earlier, the HPI model would be used for every MAP value. This application would plea for a definition of non-hypotension as any MAP > 65 mmHg. However, I chose to exclude the range between 65 and 70 mmHg in the definition of non-hypotension. As Figure 4.18 shows, the number of FPs was extremely high in range of 65 to 70 mmHg, as the PPV becomes poor when including this MAP range in the definition of non-hypotension. In detailed evaluation of the data, the excessive amount of FPs appeared to originate from long time segments with a stable MAP in this range, whilst HPI values exceeded the alarm threshold. These episodes had a large influence on the results because of the nature of the FSW labelling approach. This approach labels each single prediction in the episode as a false positive, whereas the hemodynamic situation remained unchanged for a long time. A different labelling approach, similar to the FTW protocol, could label this episode as a single prediction. However, to cover this limitation of the FSW protocol for episodes in the range of 65 to 70 mmHg, non-hypotension was defined as a MAP > 70 mmHg.

Lastly, labels were not corrected for dependent variables. In the primary protocol, each HPI value was labelled as an independent prediction. The reason behind this is that each HPI prediction is an independent calculation of the algorithm based on the data of the previous 20 seconds. However, within a series of predictions the patient is a dependent variable. So, individual predictions are not independent. If a hypotensive episode is correctly predicted in a patient, the odds are expected to be higher for the next hypotensive episode in this patient to be detected correctly as well and vice versa. This means that patients with relatively large representation of either hypotension or non-hypotension could bias the results. For example, patients with SAH introduce bias to the data as the incidence of hypotension is low due to a different targeted blood pressure, as discussed previously. To monitor the introduction of bias by certain patient categories, subgroup analysis was also performed. In addition, a washout period limits the introduction bias by dependent predictions of subsequent hypotensive episodes. Other corrections could make the labelling and calculation of performance metrics more com-

plex. A potential pitfall could be that complex validation protocols could make it difficult to comprehend the results.

5.5. Implication by study results

Validation of a predictive algorithm is an essential step in the product life-cycle. Validation serves as a pre-clinical quality check between algorithm development and studies on clinical benefit.[26,55] Especially external validation shows the expected model performance upon implementation.[48] The HPI model was developed by Edwards LifeSciences and received FDA approval. However, responsibility for clinical use of the algorithm remains with the clinician.[26]

The decision on clinical implementation of the HPI on the intensive care is supported by these validation results. However, the clinical benefit should be evaluated in clinical trials. In addition, the validation results also give direction to further research on alarm threshold selection and direction to further research on the subgroups which would benefit most from additional hemodynamic monitoring by the HPI model.

5.6. Recommendations

5.6.1. Validation protocol use

The three conceptually different protocols have different strengths and limitations. The use of each protocol can be justified, but only by using different research questions. My interpretation on advantages and justification of the use of each protocol is described below.

The BW protocol only provide an initial indication whether hypotension can be predicted by the HPI model. By using backward labelling, the protocol does not provide information on what outcome can be expected upon a particular HPI value or alarm on impending hypotension. Therefore, generalizability to clinical practice is limited. In addition, the amount of FPs by the BW protocol is tempered, as only hemodynamically stable periods are selected for labelling of non-hypotension. Time windows were only labelled as non-hypotension if all MAP values within a time window showed a MAP > 75 mmHg. So, performance metrics from the BW protocol are less valuable for interpretation of HPI predictions made during moments with glooming hemodynamic instability. In clinical practice, predictions during this twilight zone would affect treatment selection. Therefore, clinical generalizability of results from the BW protocol are hampered.

The FTW protocol mimics a clinical treatment protocol using the HPI. Only the onsets of alarms are labelled, as they indicate the initiation of pro-active treatment. This feature strengthens the generalizability of the results. However, important to note is that an element was added to the definition of an alarm, which may have clinical consequences. In the FTW protocol, a minimal duration of one minute of HPI values above alarm threshold was added to the alarm definition. This constraint raises alarm criteria and reasonably reduces FPs. This would only be justified as long as the clinical protocol of HPI use would also incorporate this minimum duration. The downside of this constraint is the reduction of timeliness of the eventual alarm. In addition, the FTW protocol does not consider that user behaviour may change when HPI decreases to sub-threshold levels after the initial alarm, which could make the clinical cancel the pro-active treatment. In the end, the advantage of the FTW protocol is the simulation

of clinical implementation. Therefore the FTW could best be used for pro-active treatment protocol evaluation.

The FSW protocol displays the performance of individual predictions and proves a detailed insight in predictive performance of HPI values. Thereby, it may provide the most realistic performance values, given that a clinician interprets individual HPI values. However, the results of the FSW are less informative on correctness of hypothetically initiated pro-active treatment, in contrast to the FTW protocol.

5.6.2. Threshold selection

What HPI threshold should be used in clinical practice on the ICU? This major question cannot be answered by statistical analysis alone. Performance metrics, as used in validation of the HPI, are ”just statistical abstractions and not yet informative about clinical value”, as described by Vickers *et al.* [57]. Eventual threshold selection should be founded on the potential harm and benefits of interventions up HPI alarm. So only the quantified potential harm and benefits of each label in combination with these validation result could lead to a calculated clinical optimum on HPI alarm threshold.

The potential harm and benefits of pro-active treatment on an alarm, whether true or false, remain to be estimated for ICU patients in future clinical studies. An ongoing clinical study at the AmsterdamUMC evaluates the effect of pro-active treatment upon an HPI value above 75. For surgical patients, Wijnberge *et al.* [59] already showed that pro-active treatment upon an HPI value above 85 reduced hypotension occurrence and severity. Pro-active treatment did not result in a statistically significant difference in cumulative dose of medication nor fluids given during surgery.[59] However, an other study reported non-significant results reduction of intraoperative hypotension, using an alarm threshold of 85.[34]

The ideal threshold is yet to be determined. For now, expert opinion, pilot studies and an HPI value of 85 as only alarm threshold approved by the FDA determine the thresholds used in research. The following considerations are recommended to be included in threshold selection for clinical use of the HPI. Firstly, the potential harm and benefits per label could be estimated using expert opinion and scientific literature, as mentioned above. A cost-benefit analysis can be performed within the domains of time, finance and patient health. A possible method to evaluate the balance between harm and benefits to the health of the patient is the Net Benefit analysis. The Net Benefit analysis is a decision analytic measure that brings potential harm to the benefits to the same scale by using an exchange rate based in clinical judgement.Vickers *et al.* [57]

Secondly, the setting in which the HPI is used also determines the desired threshold. The HPI model is certified to be applied to both surgical patients and ICU patients. As in ICU patients the changes in hemodynamic states are expected to be more gradual than in surgical patients, the expected time-to-hypotension for an alarm is longer in ICU patients.[60] This is supported by the median [IQR] time-to-hypotension, with an alarm threshold of 75, observed in ICU patients and surgical patients of 4.3 [1.6 to 11] min and 1.3 [0.7 to 4.3],[59] respectively. The extra time-to-hypotension for ICU patients could be exchanged for a reduction in false positives by increasing the alarm threshold. However, the optimal time-to-hypotension could differ per setting, as the time from alarm to initiation of treatment also differs between the two settings. During surgery, a dedicated clinician monitors the patient at all times whereas an ICU patient is monitored by a nurse that has multiple tasks or even multiple patients to look

after. Thus, a longer time-to-hypotension of an alarm may be desired in the ICU setting than in a surgical setting.

The last recommendation for alarm threshold selection is to take the effect of HPI threshold on usage behaviour into consideration. Change in behaviour of end-users is a great determinant of the impact of an innovation. To estimate difference in acceptance of the HPI algorithm for different alarm threshold conditions, the extended version of the Technology Acceptance Model (TAM2) could be used as a foundation. The TAM2 theory by Venkatesh & Davis [56] states that usage behaviour of a technology is eventually determined by both 'Perceived Usefulness' and 'Perceived Ease of Use'. The 'Perceived Usefulness' will be influenced by the HPI threshold that will be selected for standard clinical use. 'Perceived Usefulness' is theoretically affected by several factors, of which the following were identified that could be relevant in alarm threshold selection:

Both the 'Result Demonstrability' of the HPI model, i.e. the acquaintance of the user with results of using the innovation, and the 'Perceived Importance' of hypotension prevention contribute to the successful implementation of the HPI algorithm.[15,56] Facing the validity and the evidence base by end-users of an innovation is reported as an important facilitator of successful implementation.[12] A difference in attitude towards a MAP below 65 mmHg was observed between nurses in anaesthesiology and the nurses in the ICU. My personal impression was that a MAP target of > 65 mmHg was more strictly followed during surgery than during ICU admission. Albeit a single observation, a clear presentation of the evidence base behind the HPI and the presentation of results for multiple thresholds could aid the implementation of the HPI algorithm on the ICU.

Lastly, the increase of false positives by using a lower threshold value of the HPI could lead to alarm fatigue. Alarm fatigue by false positive alarms leads to a lower response rate on alarms on the ICU.[7] However, alarms are in integral part of care provided in ICUs. The further advancements in use of technology on the ICU will undoubtedly lead to an increase in alarms as well.[30] Therefore the threshold should be selected carefully to avoid alarm fatigue, loss of perceived "Output Quality" and reduced "Perceived Usefulness" of the HPI model, as defined in the TAM2[56].

5.6.3. Performance metric selection

I encourage the use of PPV and PR curves in reports on performance of hypotension prediction models. Currently, more weight is placed on specificity than on PPV in validation (Appendix J). Specificity is a good objective metric to compare different models that are tested on different data sets. The calculation of specificity is not influenced by the ratio between hypotension and non-hypotension in the data set. Therefore, it is an objective metric for comparison between models.

However, for demonstration of the clinical applicability of a predictive model, I would suggest the use of PPV instead of specificity. A hypotension prediction model would be used as a bedside warning system. As long as only alarms will affect the clinical behaviour and no-alarms will not, the user is interested in the probability of impending hypotension upon an alarm, i.e. PPV. A user would not be interested in the probability on a no-alarm given no hypotension would occur, i.e. specificity. So, clinical usefulness could better be illustrated using PPV and PR curves, rather than specificity and ROC curves. This is also applicable to threshold selection.

5.6.4. Future validation protocol options

More validation approaches and data labelling methods were identified during the project. Unfortunately, their evaluation was outside the time scope of this thesis. Therefore, the following approaches are recommended for future research on HPI or other predictive alarms using continuous data.

In alarm definition, different thresholds could be used to turn the alarm 'on' and to turn the alarm 'off'. For example, the HPI model would alarm the clinician when a HPI values exceeds 85, but would only silence this alarm when a HPI value decreases below 75. As the timeliness assessment shows, sensitivity was approximately 45% already at 20 minutes prior to onset of hypotension, with an alarm threshold at 85. The two time-to-hypotension distributions show that an early initial HPI alarm is often followed by a no-alarm towards the onset of hypotension. Different thresholds for alarm 'on' and alarm 'off' could result a more consistent and reliable alarm pattern, whilst maintaining optimal timeliness.

Another approach to obtain a more consistent alarm pattern is to use a moving average or a minimal alarm duration, as in Wijnberge *et al.* [60]. However, if this modification would be applied in clinical practice, a clinician is only alarmed when this additional condition is satisfied. Unfortunately, this reduces the timeliness and thus added value of the algorithm. To avoid reduction of timeliness, an altered clinical workflow could be designed. In this hypothetical workflow, a single HPI value above threshold would alarm the clinician to identify the cause of impending hypotension. Subsequently, the selected hemodynamic intervention would already be prepared. The intervention would only be initiated once the minimal duration condition is satisfied and a additional alarm would ring. This hypothetical workflow may not be feasible as it induces a large amount of extra work-load due to many FPs. But it shows a possible integration of HPI value behaviour to a pro-active treatment protocol.

A different validation approach could include three categories to classify HPI values: 'no-alarm or safe' – no impending hypotension expected, 'Caution' – low probability on impending hypotension or an expected long time-to-hypotension, and 'Alarm' – alarm on impending hypotension. Thereby, the validation protocol simulates the nuance that continuous value implies when displayed on the monitor. However, the 'caution' class should only be added to the validation protocol if the 'caution' zone has a clinical implication. An example is already included in a clinical trial on the effect of HPI. This study protocol recommends the clinician to already identify the cause of probable impending hemodynamic instability at HPI levels between 50 and 75.

A weighted scoring system could be applied to predictions on hypotension, i.e. TP and FN labels. TP that are rapidly followed by hypotension would earn less points than TPs that predict the onset of hypotension far in advance. The opposite could apply for prediction on non-hypotension: an FN prediction far before onset of hypotension would be punished less than an FN prediction right before onset of hypotension. The point scoring system should only be used to compare different predictive models. A single score does not give a clear indication of absolute clinical usefulness.

Furthermore, the effect of certain protocol elements (e.g. prediction window size) could be displayed via 'iso-threshold' PR and ROC curve. Currently, the effect of protocol elements was displayed via PR or ROC curves over the full range of HPI values(1, 2, .., 100), but for a small number of values of the protocol element (e.g. 10, 15 or 20 minutes). Thereby, the effect of the varying value could be evaluated for a curve as a whole, but not for single HPI

values, because the ability to discriminate colours is limited. The proposed alternative is to plot ‘iso-threshold lines’. For example, in a PR curve, the effect of prediction window size would be illustrated by plotting the a line between points calculated for the *full range of protocol element values* (e.g. 1, 2, .., 20 minutes), but using *a single HPI threshold*. By repeating this for multiple thresholds (e.g. 5, 10, .., 100) several quasi-parallel lines would indicate the effect of a changing variable value.

The last proposal for future exploratory analysis is to evaluate the HPI performance while excluding all predictions made with a MAP < 70 mmHg. Results of this validation option may form an solution to the discussion whether the definition of non-hypotension should include MAP values between 65 and 75. Results would show a performance without any artificial polarisation. The obvious limitation is that the results cannot be extrapolated to any prediction with a MAP < 70 mmHg.

Conclusion

Conclusion

This thesis demonstrates the importance of validation methodology of the Hypotension Prediction Index on the ICU population. Labelling of individual predictions via the FSW protocol resulted in a lower PPV and sensitivity of the HPI model compared to previously reported HPI performance, in which a time windowed labelling method was used. The HPI model showed poor performance for patients with a subarachnoid haemorrhage, but a better trade-off between sensitivity and PPV for patients with a cardiogenic shock. The inclusion of mean blood pressure values in the range between 65 and 75 mmHg in the definition of non-hypotension dominantly reduced PPV. Overall, the results show the ability of the HPI to predict hemodynamic instability in ICU patients. Therefore, validation results support the introduction of the HPI to the ICU for clinical use. However, the optimal alarm threshold and clinical benefit remain to be evaluated in future clinical studies.

References

1. Antonelli, M. *et al.* Hemodynamic monitoring in shock and implications for management: International Consensus Conference, Paris, France, 27–28 April 2006. *Intensive Care Medicine* **33,** 575–590. doi:10.1007/s00134-007-0531-4 (Apr. 2007).
2. Badin, J. *et al.* Relation between mean arterial pressure and renal function in the early phase of shock: a prospective, explorative cohort study. *Critical Care 2011 15:3* **15,** 1–12. doi:10.1186/CC10253 (June 2011).
3. Bagshaw, S. M. *et al.* A Multi-Center Evaluation of Early Acute Kidney Injury in Critically Ill Trauma Patients. *Renal Failure* **30,** 581–589. doi:10.1080/08860220802134649 (Jan. 2008).
4. Bijker, J. B. *et al. Incidence of Intraoperative Hypotension as a Function of the Chosen Definition Literature Definitions Applied to a Retrospective Cohort Using Automated Data Collection* tech. rep. (2007), 213–233.
5. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30,** 1145–1159. doi:10.1016/S0031-3203(96)00142-2 (July 1997).
6. Chacon, S. & Straub, B. *Pro Git* doi:10.1007/978-1-4842-0076-6 (Apress, Berkeley, CA, 2014).
7. Cho, O. M. *et al.* Clinical Alarms in Intensive Care Units: Perceived Obstacles of Alarm Management and Alarm Fatigue in Nurses. *Healthcare Informatics Research* **22,** 46. doi:10.4258/HIR.2016.22.1.46 (Jan. 2016).
8. Donald, R. *et al.* Forewarning of hypotensive events using a Bayesian artificial neural network in neurocritical care. *Journal of Clinical Monitoring and Computing* **33,** 39–51. doi:10.1007/s10877-018-0139-y (Feb. 2019).
9. Eshelman, L. J. *et al.* Development and evaluation of predictive alerts for hemodynamic instability in ICU patients. *AMIA Annual Symposium proceedings* **2008,** 379–83 (Nov. 2008).
10. Finnerty, F. A. *et al.* Cerebral hemodynamics during cerebral ischemia induced by acute hypotension. *The Journal of clinical investigation* **33,** 1227–1232. doi:10.1172/JCI102997 (Sept. 1954).
11. Fitch, W. *et al.* Effects of decreasing arterial blood pressure on cerebral blood flow in the baboon. Influence of the sympathetic nervous system. *Circulation Research* **37,** 550–557. doi:10.1161/01.RES.37.5.550 (1975).
12. Francis, J. J. *et al.* Selective decontamination of the digestive tract in critically ill patients treated in intensive care units: a mixed-methods feasibility study (the SuDDICU study). *Health Technology Assessment* **18,** 1–170. doi:10.3310/HTA18250 (Apr. 2014).
13. French, C. *Data Processing and Information Technology* 10th (ed Cengage Learning EMEA) (Thomson Learning, 1996).

14. Gamper, G. *et al. Vasopressors for hypotensive shock* Feb. 2016. doi:10.1002/14651858.CD003709.pub4.

15. Geerligs, L. *et al.* Hospital-based interventions: a systematic review of staff-reported barriers and facilitators to implementation processes. *Implementation Science 2018 13:1* **13,** 1–17. doi:10.1186/S13012-018-0726-9 (Feb. 2018).

16. Grand, J. *et al.* Arterial blood pressure during targeted temperature management after out-of-hospital cardiac arrest and association with brain injury and long-term cognitive function. *European Heart Journal. Acute Cardiovascular Care* **9,** S122–S130. doi:10.1177/2048872619860804 (Nov. 2020).

17. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585,** 357–362. doi:10.1038/s41586-020-2649-2 (Sept. 2020).

18. Harten, J. & Kinsella, J. *Perioperative optimisation* Feb. 2004. doi:10.1177/0036933004049001 02.

19. Hatib, F. *et al.* Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. *Anesthesiology* **129,** 663–674. doi:10.1097/ALN.0000000000002300 (Oct. 2018).

20. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21,** 1263–1284. doi:10.1109/TKDE.2008.239 (Sept. 2009).

21. He, H. & Ma, Y. *Imbalanced Learning* (eds He, H. & Ma, Y.) 1–210. doi:10.1002/9781118646106 (Wiley, June 2013).

22. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in science \& engineering* **9,** 90–95 (2007).

23. Justice, A. C. *et al.* Assessing the generalizability of prognostic information. *Annals of Internal Medicine* **130,** 515–524. doi:10.7326/0003-4819-130-6-199903160-00016 (Mar. 1999).

24. Kato, R. & Pinsky, M. R. Personalizing blood pressure management in septic shock. *Annals of Intensive Care* **5,** 41. doi:10.1186/s13613-015-0085-5 (2015).

25. Kendale, S. *et al.* Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. *Anesthesiology* **129,** 675–688. doi:10.1097/ALN.0000000000002374 (Oct. 2018).

26. Komorowski, M. Clinical management of sepsis can be improved by artificial intelligence: yes. *Intensive Care Medicine 2019 46:2* **46,** 375–377. doi:10.1007/S00134-019-05898-2 (Dec. 2019).

27. Kubben, P. *et al. Fundamentals of Clinical Data Science* (eds Kubben, P. *et al.*) 1–219. doi:10.1007/978-3-319-99713-1 (Springer International Publishing, Cham, Dec. 2019).

28. Lehman, L.-W. *et al.* Hypotension as a Risk Factor for Acute Kidney Injury in ICU Patients. eng. *Computing in cardiology* **37,** 1095–1098 (2010).

29. Leone, M. *et al. Optimizing mean arterial pressure in septic shock: A critical reappraisal of the literature* Dec. 2015. doi:10.1186/s13054-015-0794-z.

30. Lewandowska, K. *et al.* Impact of Alarm Fatigue on the Work of Nurses in an Intensive Care Environment—A Systematic Review. *International Journal of Environmental Research and Public Health* **17,** 1–14. doi:10.3390/IJERPH17228409 (Nov. 2020).

31. Luo, W. *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research* **18.** doi:10.2196/jmir.5870 (Dec. 2016).

32. MATLAB. *version 7.10.0 (R2018b)* (The MathWorks Inc., Natick, Massachusetts, 2018).

33. Maheshwari, K. *et al.* The relationship between ICU hypotension and in-hospital mortality and morbidity in septic patients. *Intensive Care Medicine* **44,** 857–867. doi:10.1007/s00134-018-5218-5 (June 2018).

34. Maheshwari, K. *et al.* Hypotension Prediction Index for Prevention of Hypotension during Moderate- To High-risk Noncardiac Surgery: A Pilot Randomized Trial. *Anesthesiology* **133,** 1214–1222. doi:10.1097/ALN.0000000000003557 (2020).

35. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405,** 442–451. doi:10.1016/0005-2795(75)90109-9 (Oct. 1975).

36. McKinney, W. & others. *Data structures for statistical computing in python* in *Proceedings of the 9th Python in Science Conference* **445** (2010), 51–56.

37. Moghadam, M. C. *et al.* A machine-learning approach to predicting hypotensive events in ICU settings. eng. *Computers in Biology and Medicine* **118,** 103626. doi:10.1016/j.compbiomed.2020.103626 (Mar. 2020).

38. Parker, T. *et al.* Optimising organ perfusion in the high-risk surgical and critical care patient: a narrative review. *British Journal of Anaesthesia* **123,** 170–176. doi:10.1016/J.BJA.2019.03.027 (Aug. 2019).

39. Pathinarupothi, R. K. *et al. Deriving High Performance Alerts from Reduced Sensor Data for Timely Intervention in Acute Hypotensive Episodes* in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* **2018-July** (IEEE, July 2018), 3260–3263. doi:10.1109/EMBC.2018.8512945.

40. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12,** 2825–2830 (2011).

41. Poukkanen, M. *et al.* Hemodynamic variables and progression of acute kidney injury in critically ill patients with severe sepsis: data from the prospective observational FINNAKI study. *Critical Care 2013 17:6* **17,** 1–11. doi:10.1186/CC13161 (Dec. 2013).

42. Rinkel GJE, e. a. *Richtlijn Subarachnoïdale Bloeding: Bloeddruk bij een SAB* 2013.

43. Saugel, B. *et al. Predicting hypotension in perioperative and intensive care medicine* June 2019. doi:10.1016/j.bpa.2019.04.001.

44. Schenk, J. *et al.* Definition and incidence of hypotension in intensive care unit patients, an international survey of the European Society of Intensive Care Medicine. *Journal of Critical Care* **65,** 142–148. doi:10.1016/J.JCRC.2021.05.023 (Oct. 2021).

45. Seymour, C. W. *et al.* Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA - Journal of the American Medical Association* **315,** 762–774. doi:10.1001/jama.2016.0288 (Feb. 2016).

46. Shoemaker, W. C. *et al.* Role of oxygen debt in the development of organ failure sepsis, and death in high-risk surgical patients. *Chest* **102,** 208–215. doi:10.1378/chest.102.1.208 (July 1992).

47. Smischney, N. J. *et al.* Postoperative hypotension in patients discharged to the intensive care unit after non-cardiac surgery is associated with adverse clinical outcomes. *Critical Care 2020 24:1* **24,** 1–12. doi:10.1186/S13054-020-03412-5 (Dec. 2020).

48. Steyerberg, E. *Clinical Prediction Models* doi:10.1007/978-0-387-77244-8 (Springer New York, New York, NY, 2009).

49. Steyerberg, E. W. & Vergouwe, Y. *Towards better clinical prediction models: Seven steps for development and an ABCD for validation* Aug. 2014. doi:10.1093/eurheartj/ehu207.

50. Steyerberg, E. W. *et al. Assessing the performance of prediction models: A framework for traditional and novel measures* Jan. 2010. doi:10.1097/EDE.0b013e3181c30fb2.

51. Strandgaard, S. Autoregulation of cerebral blood flow in hypertensive patients. The modifying influence of prolonged antihypertensive treatment on the tolerance to acute, drug induced hypotension. *Circulation* **53,** 720–727. doi:10.1161/01.CIR.53.4.720 (1976).

52. Sun, L. Y. *et al.* Association of intraoperative hypotension with acute kidney injury after elective noncardiac surgery. *Anesthesiology* **123,** 515–523. doi:10.1097/ALN.0000000000000765 (Sept. 2015).

53. Trzeciak, S. *et al.* Significance of arterial hypotension after resuscitation from cardiac arrest*. *Critical Care Medicine* **37,** 2895–2903. doi:10.1097/CCM.0b013e3181b01d8c (Nov. 2009).

54. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

55. Van der Ven, W. H. *et al.* One of the first validations of an artificial intelligence algorithm for clinical use: The impact on intraoperative hypotension prediction and clinical decision-making. eng. *Surgery (United States).* doi:10.1016/j.surg.2020.09.041 (Dec. 2020).

56. Venkatesh, V. & Davis, F. D. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* **46,** 186–204. doi:10.1287/mnsc.46.2.186.11926 (Feb. 2000).

57. Vickers, A. J. *et al.* Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352.** doi:10.1136/BMJ.I6 (Jan. 2016).

58. Vincent, J.-L. *et al.* Mean arterial pressure and mortality in patients with distributive shock: a retrospective analysis of the MIMIC-III database. *Annals of Intensive Care* **8,** 107. doi:10.1186/s13613-018-0448-9 (Dec. 2018).

59. Wijnberge, M. *et al.* Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension during Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. *JAMA - Journal of the American Medical Association* **323,** 1052–1060. doi:10.1001/jama.2020.0592 (Mar. 2020).

60. Wijnberge, M. *et al.* Clinical performance of a machine-learning algorithm to predict intraoperative hypotension with noninvasive arterial pressure waveforms: A cohort study. *European journal of anaesthesiology* **38,** 609–615. doi:10.1097/EJA.0000000000001521 (June 2021).

Apendices

HPI design

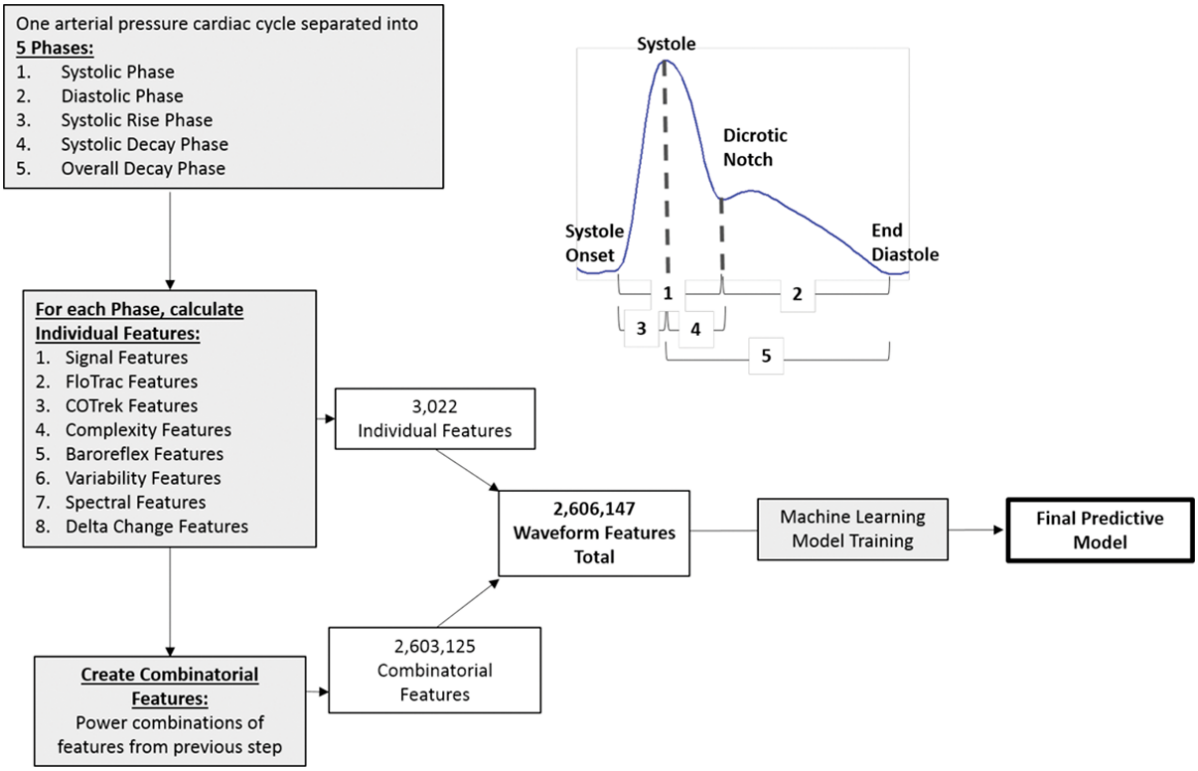


Figure A.1: A High-level overview of the development of the Hypotension Prediction Index, based on arterial blood pressure waveforms. Model development is described in the original article by Hatib *et al.* [19]. The image was reprinted from Hatib *et al.* [19], with permission from Elsevier.

Exclusion per label type

Table B.1: Data points excluded as erroneuos data per label, for FSW protocol.

Labels	Total	Excluded	% Excluded
TP	22364	1018	4,551959
FP	31256	206	0,659073
TN	476467	5905	1,23933
FN	16519	1946	11,78037

Forward sliding window validation: Overview thresholds

Table C.1: Performance according to forward sliding window validation protocol for all thresholds as multiples of five and statistically optimal thresholds. Min. diff. Se and Sp: The threshold that minimalizes the difference between sensitivity and specificity.

HPI threshold	Se	Sp	PPV	NPV	Threshold method
5	0.995	0.1014	0.0735	0.9965	
10	0.9901	0.1744	0.0791	0.996	
15	0.981	0.2425	0.0849	0.9944	
20	0.9697	0.3065	0.091	0.993	
25	0.9575	0.3711	0.0983	0.9919	
30	0.9396	0.4415	0.1075	0.9903	
35	0.9179	0.5225	0.121	0.9889	
40	0.8929	0.6015	0.1383	0.9874	
45	0.8705	0.6571	0.1538	0.9861	
50	0.8475	0.6986	0.1676	0.9846	
55	0.8225	0.7369	0.1829	0.983	
60	0.7918	0.7759	0.2019	0.9812	
65	0.7549	0.8165	0.2276	0.979	
70	0.72	0.8508	0.2569	0.977	
75	0.683	0.8815	0.2922	0.9749	
80	0.6419	0.9099	0.3379	0.9726	
85	0.5943	0.9381	0.4074	0.97	
90	0.5282	0.9665	0.5303	0.9662	
95	0.3954	0.9943	0.832	0.9583	
60	0.7918	0.7759	0.2019	0.9812	Min. diff. Se and Sp
64	0.7624	0.8089	0.2222	0.9794	Youden
94	0.438	0.9894	0.7466	0.9609	F1 (score: 0.55)

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value;

Forward tumbling window validation: Overview thresholds

Table D.1: Performance according to forward tumbling window validation protocol for all thresholds as multiples of five and statistically optimal thresholds. Min. diff. Se and Sp: The threshold that minimalizes the difference between sensitivity and specificity.

HPI threshold	Se	Sp	PPV	NPV	
5	1	0.11	0.6713	1	
10	1	0.1676	0.6757	1	
15	1	0.23	0.683	1	
20	1	0.2821	0.6913	1	
25	1	0.3323	0.6989	1	
30	1	0.3857	0.7082	1	
35	1	0.4494	0.7223	1	
40	1	0.5069	0.7381	1	
45	1	0.54	0.748	1	
50	1	0.5721	0.7582	1	
55	0.9999	0.5991	0.7667	0.9997	
60	0.9999	0.6271	0.7772	0.9997	
65	0.9999	0.6584	0.7878	0.9998	
70	0.9994	0.688	0.7998	0.9989	
75	0.9991	0.7136	0.8104	0.9985	
80	0.9981	0.7341	0.8172	0.9969	
85	0.9972	0.7625	0.8304	0.9957	
90	0.9935	0.8054	0.8508	0.9911	
95	0.9764	0.8963	0.8979	0.976	
95	0.9764	0.8963	0.8979	0.976	Min. diff. Se and Sp
95	0.9764	0.8963	0.8979	0.976	Youden
95	0.9764	0.8963	0.8979	0.976	F1 (score: 0.94)

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value.

Backward validation: Overview thresholds

Table E.1: Performance according to backward validation protocol for all thresholds as multiples of five and statistically optimal thresholds. Min. diff. Se and Sp: The threshold that minimalizes the difference between sensitivity and specificity.

t	HPI threshold	Se	Sp	PPV	NPV	Threshold method
5	5	0.999	0.157	0.648	0.989	
5	10	0.997	0.269	0.68	0.984	
5	15	0.995	0.379	0.714	0.98	
5	20	0.993	0.478	0.748	0.977	
5	25	0.989	0.578	0.785	0.971	
5	30	0.984	0.678	0.826	0.964	
5	35	0.978	0.784	0.876	0.958	
5	40	0.971	0.873	0.923	0.951	
5	45	0.964	0.923	0.951	0.943	
5	50	0.957	0.951	0.968	0.935	
5	55	0.948	0.968	0.979	0.924	
5	60	0.938	0.984	0.989	0.911	
5	65	0.927	0.996	0.997	0.898	
5	70	0.913	0.998	0.999	0.881	
5	75	0.9	0.998	0.999	0.865	
5	80	0.881	0.999	0.999	0.843	
5	85	0.858	0.999	1.0	0.819	
5	90	0.825	1.0	1.0	0.786	
5	95	0.752	1.0	1.0	0.721	
5	100	0.252	1.0	1.0	0.462	
5	51	0.955	0.956	0.971	0.932	Min. diff. Se and Sp
5	64	0.931	0.995	0.997	0.902	Youden
10	5	0.998	0.158	0.646	0.979	
10	10	0.995	0.267	0.677	0.974	
10	15	0.992	0.375	0.71	0.968	
10	20	0.989	0.476	0.745	0.965	
10	25	0.984	0.578	0.782	0.959	
10	30	0.977	0.675	0.823	0.95	
10	35	0.969	0.783	0.873	0.943	
10	40	0.961	0.869	0.919	0.935	
10	45	0.952	0.912	0.944	0.925	
10	50	0.945	0.943	0.963	0.917	

Table E.1 continued from previous page

t	HPI threshold	Se	Sp	PPV	NPV	Threshold method
10	55	0.933	0.963	0.975	0.903	
10	60	0.921	0.981	0.987	0.89	
10	65	0.903	0.991	0.993	0.869	
10	70	0.889	0.997	0.998	0.853	
10	75	0.872	0.999	0.999	0.834	
10	80	0.855	0.999	0.999	0.817	
10	85	0.832	0.999	1.0	0.794	
10	90	0.792	1.0	1.0	0.757	
10	95	0.712	1.0	1.0	0.693	
10	100	0.275	1.0	1.0	0.472	
10	50	0.945	0.943	0.963	0.917	Min. diff. Se and Sp
10	59	0.925	0.978	0.985	0.894	Youden
15	5	0.998	0.151	0.643	0.983	
15	10	0.996	0.259	0.673	0.978	
15	15	0.993	0.349	0.7	0.969	
15	20	0.988	0.427	0.725	0.957	
15	25	0.982	0.5	0.75	0.948	
15	30	0.975	0.56	0.772	0.935	
15	35	0.967	0.624	0.797	0.925	
15	40	0.954	0.677	0.819	0.907	
15	45	0.944	0.708	0.832	0.892	
15	50	0.936	0.737	0.845	0.883	
15	55	0.925	0.766	0.858	0.87	
15	60	0.911	0.799	0.874	0.854	
15	65	0.894	0.831	0.89	0.837	
15	70	0.878	0.868	0.91	0.823	
15	75	0.856	0.9	0.929	0.804	
15	80	0.833	0.924	0.944	0.783	
15	85	0.807	0.944	0.957	0.762	
15	90	0.768	0.96	0.967	0.73	
15	95	0.686	0.979	0.98	0.67	
15	100	0.274	0.998	0.996	0.473	
15	71	0.875	0.873	0.913	0.82	Min. diff. Se and Sp
15	79	0.837	0.921	0.942	0.787	Youden

Abbreviations: *Se*, sensitivity; *Sp*, Specificity; *PPV*, positive predictive value; *NPV*, negative predictive value.

Additional figures on secondary analyses

F.1. Subgroup analyses

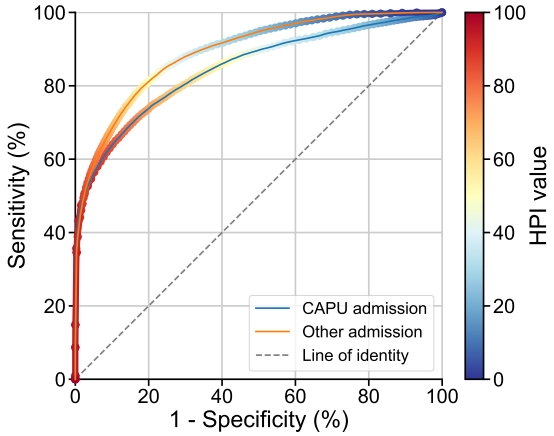


Figure F.1: ROC curve for patients admitted after cardiothoracic surgery.

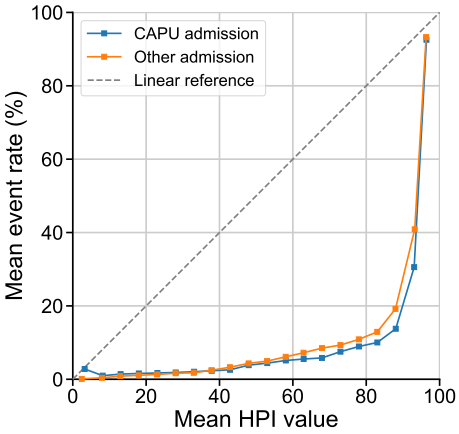


Figure F.2: Calibration curve for patients admitted after cardiothoracic surgery.

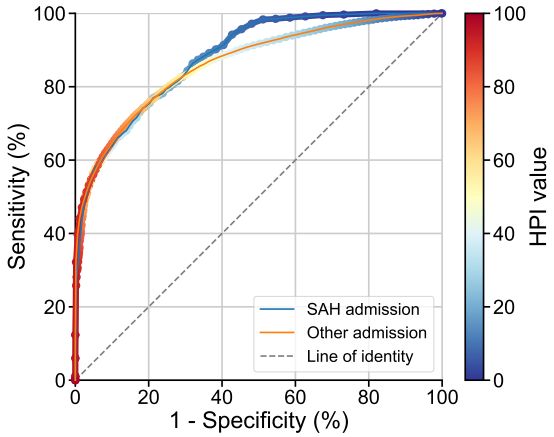


Figure F.3: ROC curve for patients with a subarachnoid haemorrhage.

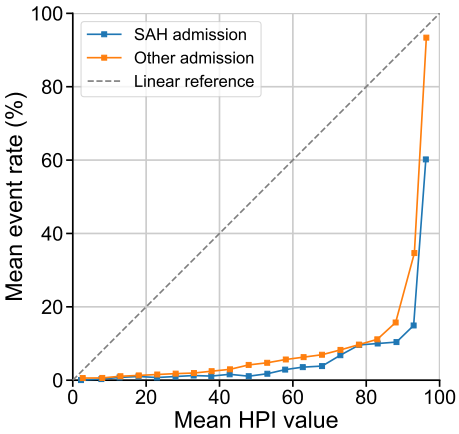


Figure F.4: Calibration curve for patients with a subarachnoid haemorrhage.

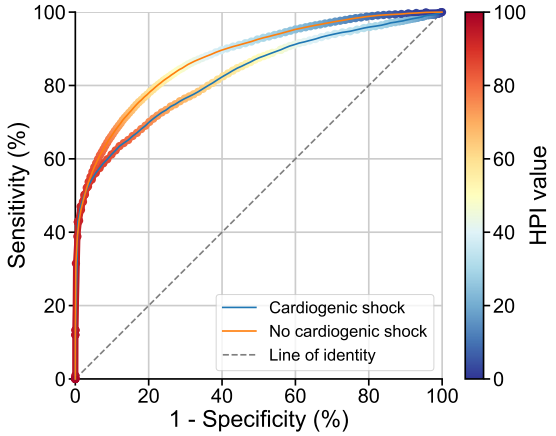


Figure F.5: ROC curve for patients with a cardiogenic shock.

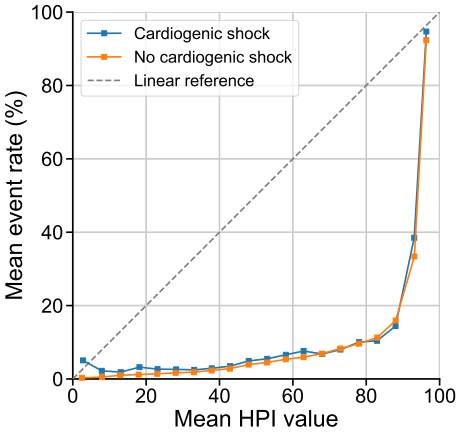


Figure F.6: Calibration curve for patients with a cardiogenic shock.

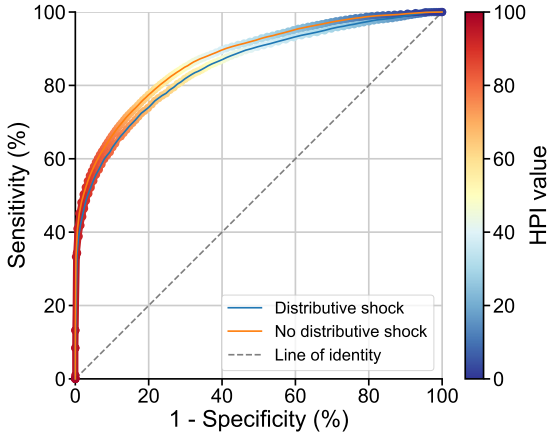


Figure F.7: ROC curve for patients with a distributive shock.

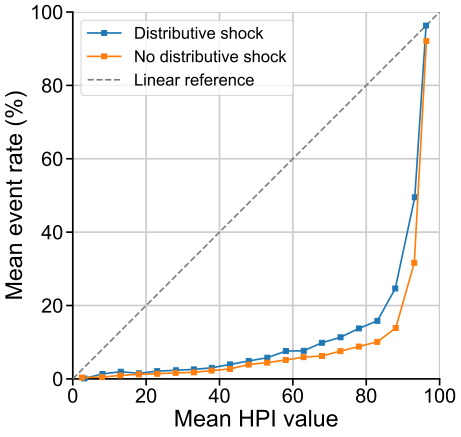


Figure F.8: Calibration curve for patients with a distributive shock.

F.2. Non-hypotension definition

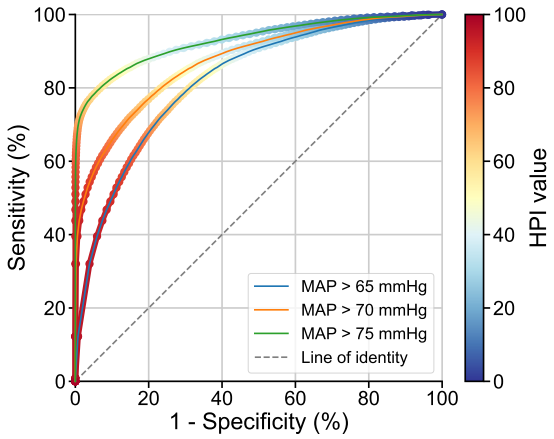


Figure F.9: ROC curves for different minimal MAP values in the definition of non-hypotension.

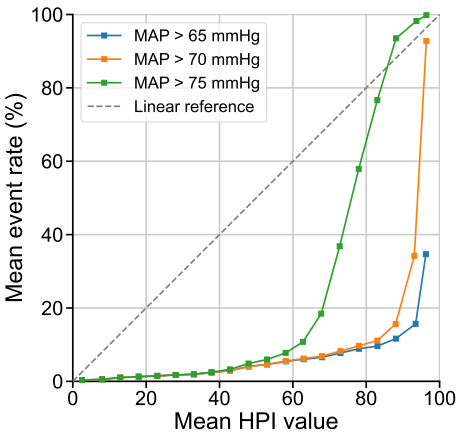


Figure F.10: Calibration curves for different minimal MAP values in the definition of non-hypotension

F.3. Prediction window duration

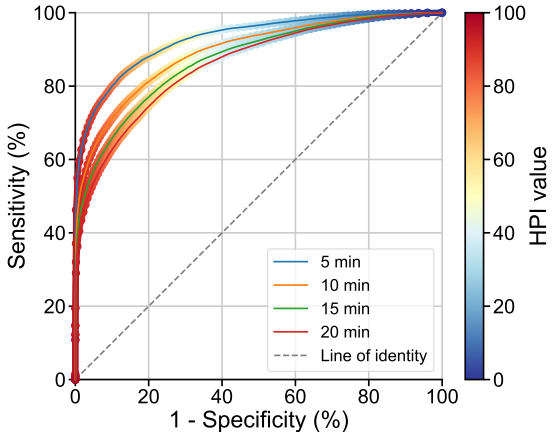


Figure F.11: ROC curves for different prediction window durations, supplemented with leading neutral buffer points to 20 minutes.

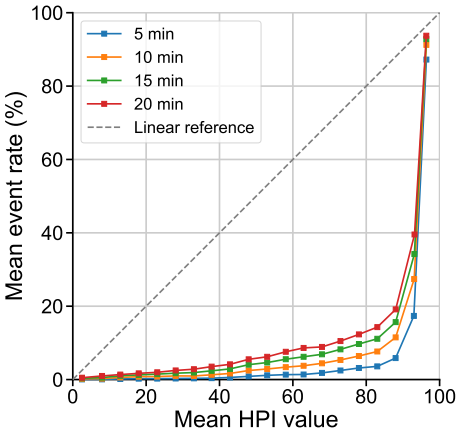


Figure F.12: Calibration curves for different prediction window durations, supplemented with leading neutral buffer points to 20 minutes.

F.4. Leading neutral buffer duration

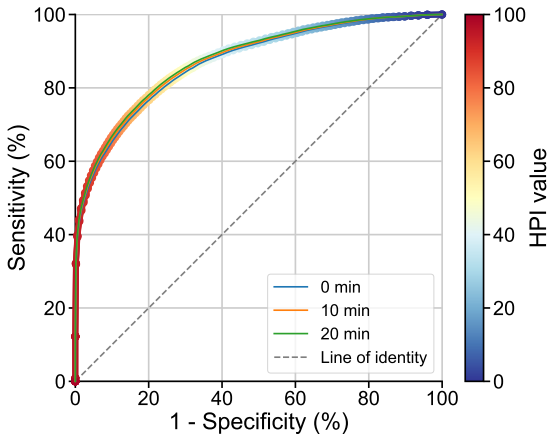


Figure F.13: ROC curves for different leading neutral buffer durations.

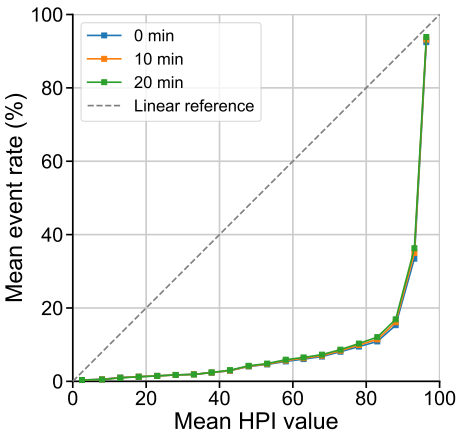


Figure F.14: Calibration curves for different leading neutral buffer durations.

F.5. Washout periods

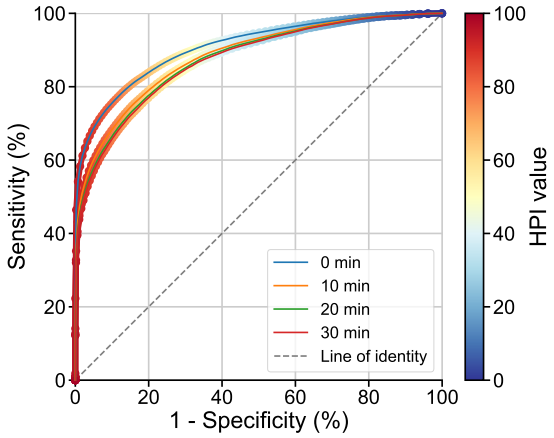


Figure F.15: ROC curves for different washout periods.

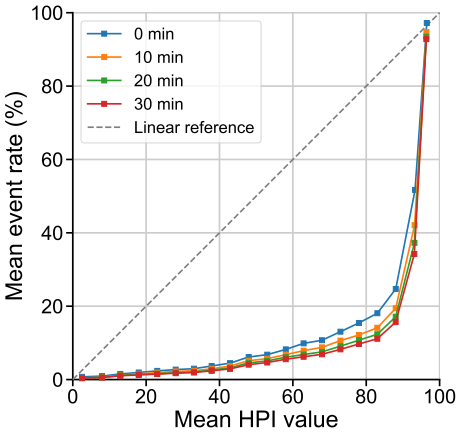


Figure F.16: Calibration curves for different washout periods.

Time-to-hypotension for different alarm thresholds

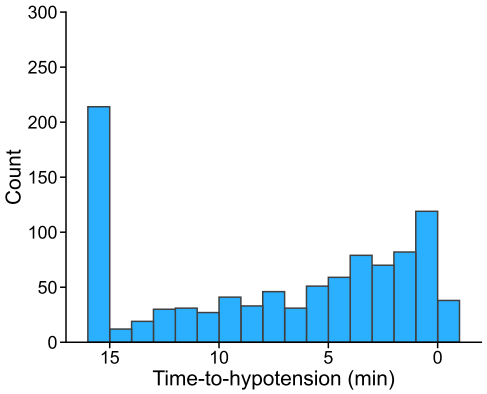


Figure G.1: Backward timeliness assessment, with alarm threshold at 65.

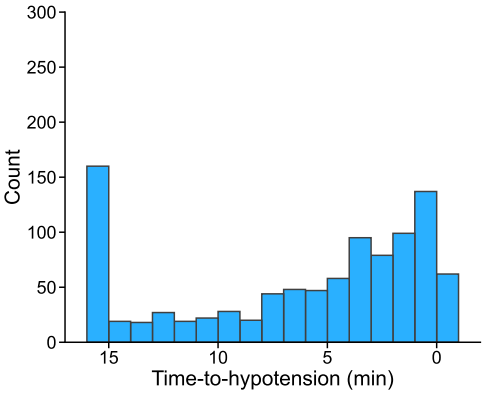


Figure G.2: Backward timeliness assessment, with alarm threshold at 75.

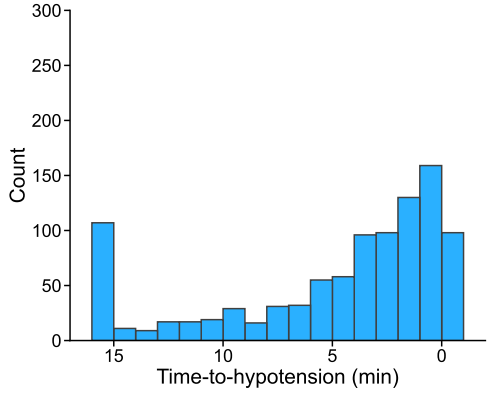


Figure G.3: Backward timeliness assessment, with alarm threshold at 85.

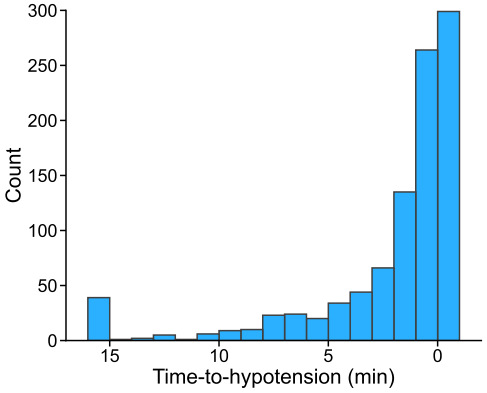


Figure G.4: Backward timeliness assessment, with alarm threshold at 95.

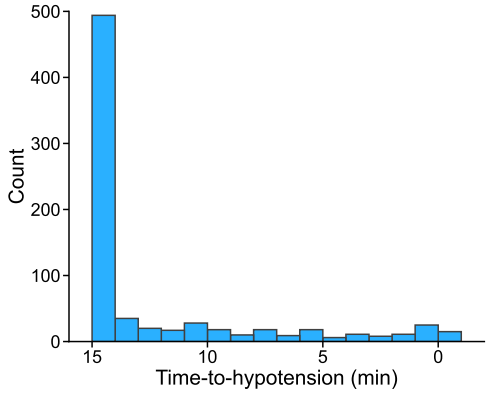


Figure G.5: Forward timeliness assessment, with alarm threshold at 65.

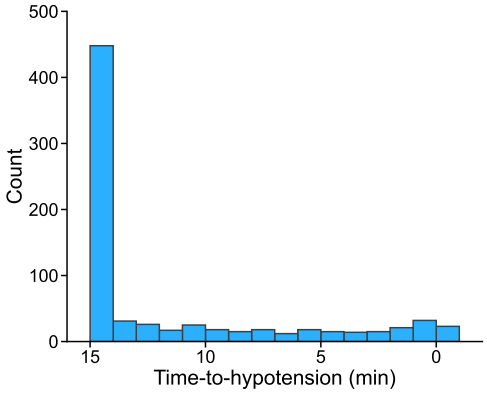


Figure G.6: Forward timeliness assessment, with alarm threshold at 75.

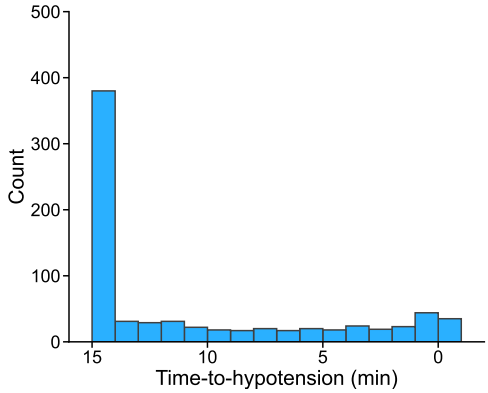


Figure G.7: Forward timeliness assessment, with alarm threshold at 85.

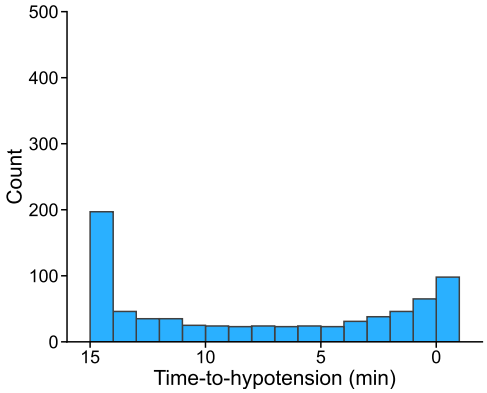


Figure G.8: Forward timeliness assessment, with alarm threshold at 95.

The effect of undersampling

PPV is the ratio of correct alarms (TP) to all alarms (TP + FP). Therefore, decreasing the number of FPs increases the PPV. One way of decreasing the number of FPs is to decrease the number of predictions on non-hypotension. Reduction of non-hypotension prediction can be achieved by 'undersampling', which involves only including a subset of the total labels for analysis.

This step was performed to illustrate the effect that undersampling would have on the reported performance metrics. In this analysis, the non-hypotension predictions were undersampled to 20% of the total number, as performed in Moghadam *et al.* [37].

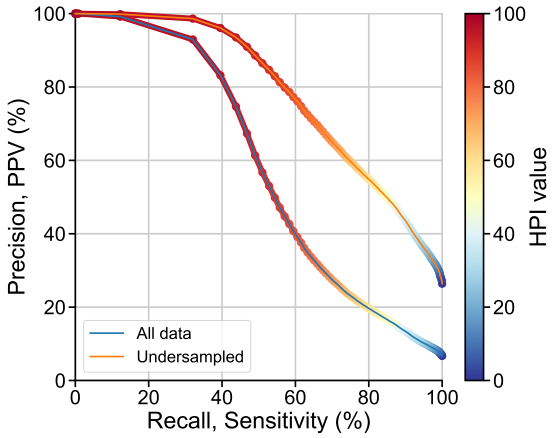


Figure H.1: PR curves for undersampled and original class distribution

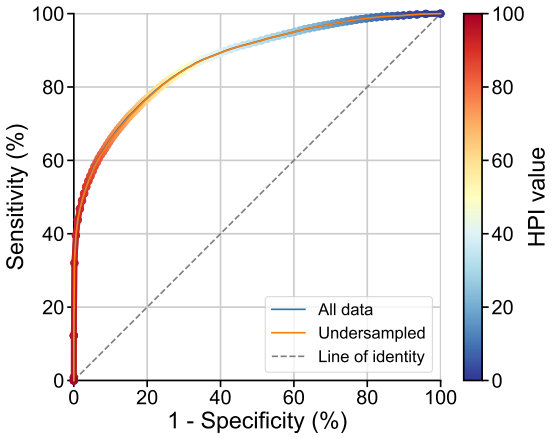


Figure H.2: ROC curves for undersampled and original class distribution

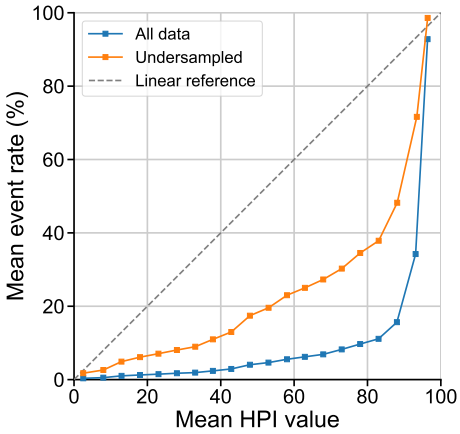


Figure H.3: Calibration curves for undersampled and original class distribution

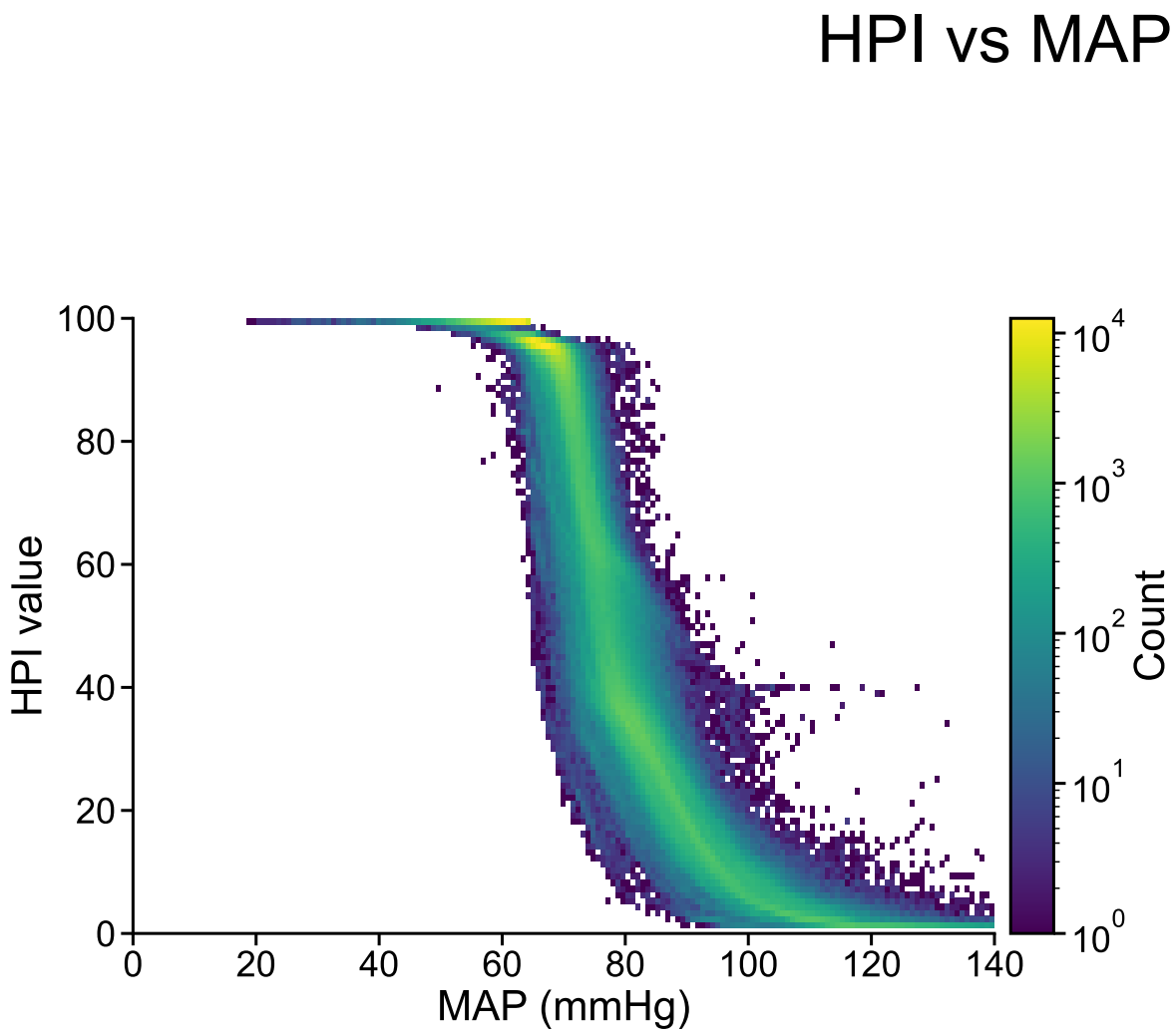


Figure I.1: HPI vs MAP: all predictions (TP, TN, FP and FN)

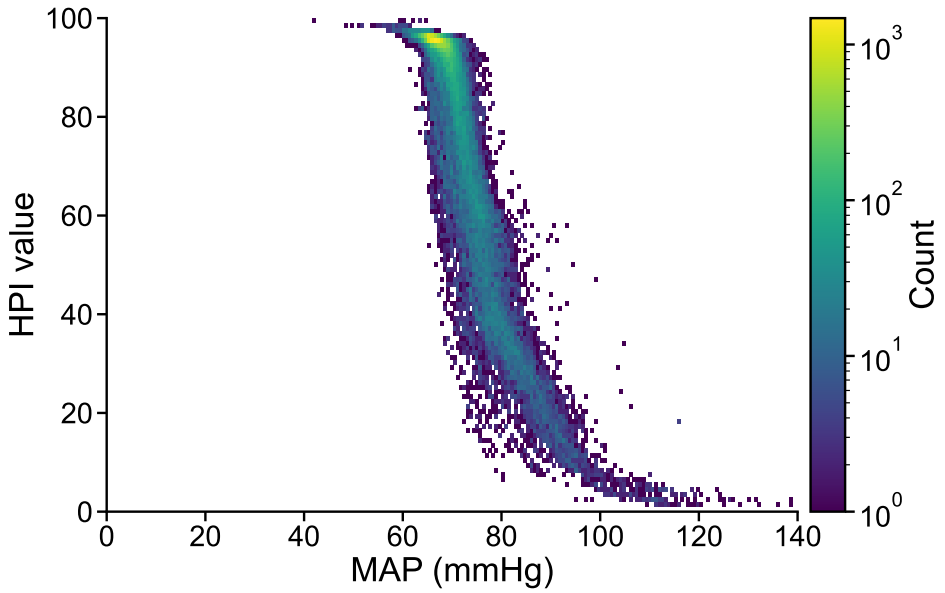


Figure I.2: HPI vs MAP: positive predictions (TN and FP)

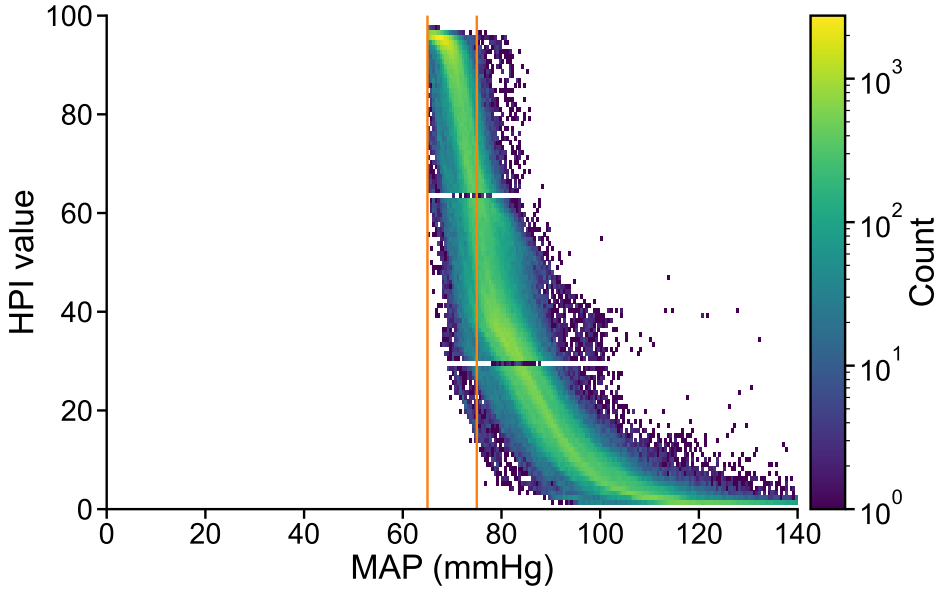
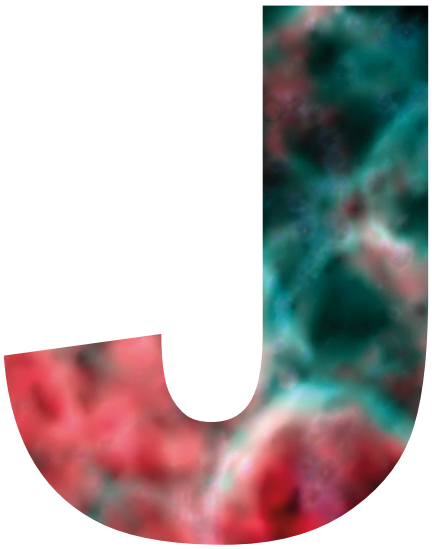


Figure I.3: HPI vs MAP: negative predictions (TN and FP)



Literature Study

Hypotension Prediction

Performance Metrics and Definitions of Hypotension Used in Hypotension Prediction Models for Intensive Care Patients

Literature Study (TM30003)

M.P. Ligtenberg BSc

TU Delft

Amsterdam UMC
Universitair Medische Centra

Hypotension Prediction

Performance Metrics and Definitions of Hypotension Used in Hypotension Prediction Models for Intensive Care Patients

by

M.P. Ligtenberg BSc

Medical Supervisor:	prof. A.P.J. Vlaar (Alexander), MD, PhD
Technical Supervisor:	E. Demirović (Emir), PhD
Daily Supervisor:	J.P. van der Ster (Björn), PhD
Institutions:	Delft University of Technology
	Amsterdam UMC, Locatie AMC
Project Duration:	January 2021 to April 2021

TU Delft

Preface

This literature study is the first part of the Master Thesis Project, the pinnacle of the master programme of Technical Medicine. The interdisciplinary joint-degree Master programme in technical medicine is offered by Delft University of Technology, together with Leiden University, Erasmus University Rotterdam and their academic medical centres. Nonetheless, I conduct my Master Thesis Project at the haemodynamic research group within the department of Intensive Care medicine of Amsterdam University Medical Centre, Location AMC.

I would like to thank my daily supervisor Björn van der Ster for our talks and direct supervision of my project. I thank my medical supervisor Alexander Vlaar for his warm invitation to the haemodynamic research group and his focus on clinical usefulness of new technology. I also want to thank my technical supervisor Emir Demorović from the faculty of Engineering, Mathematics and Computer Science of Delft University of Technology, for taking the leap into the medical domain and providing technical guidance on the project. Lastly, I want to show my appreciation to my direct colleagues for including me into their team.

M.P. Ligtenberg BSc
Rotterdam, April 2021

Contents

Abbreviation List

Abbreviation	Definition
ABP	arterial blood pressure
ACC	accuracy
AHE	acute hypotensive event
AUC	area under the curve
AUROC	area under the ROC curve
DBP	diastolic blood pressure
FN	false negative
FP	false positive
FPR	fasle postive rate: 1 - specificity
ICU	intensive care unit
IOH	intraoperative hypotension
MAP	mean arterial blood pressure
mmHg	millimetres of mercury, unit of pressure, 100 mmHg = 13,3 kPa
NPV	negative predictive value
PPV	positive predictive value
PR curve	precision-recall curve
PRAUC	area under the PR-curve
ROC curve	receiver operator characteristic curve
Recall	sensitivity
SBP	systolic blood pressure
SE	sensitivity
SP	specificity
TN	true negative
TP	true positive
TPR	true positive rate, see sensitivity

1 Introduction

Hypotension during noncardiac surgery and intensive care unit (ICU) admission due to sepsis is associated with occurrence of acute kidney injury and mortality rate.[1–5] Prevention of hypotension is also included in treatment guidelines for septic patients.[6,7] Reducing acute hypotensive event (AHE) duration or severity could lead to improved clinical outcome of patients.[2,8] Current interventions on AHE are mainly reactive. However, multiple predictive models have been designed to alarm for impending hypotension.[9–13] Such alarm enables clinicians to pro-actively start treatment and increases time for clinicians to prepare the indicated intervention. Timely intervention could reduce hypotension occurrence and severity in the intensive care setting, as already demonstrated during noncardiac surgery.[14]

Predictive computer models, or algorithms, require thorough retrospective quality assessment before clinical introduction of a hypotension. Model quality regards the degree in which predictions of the model correspond with the actual occurrence of hypotension. Correct quantification of model quality is of paramount importance in evaluating the models potentially added clinical value, as well as the potential harm on incorrect predictions.[15]

In design and validation of hypotension prediction algorithms for ICU patients, fundamental inconsistencies were found in earlier reported definitions of hypotensive events, as well as metrics used to quantify model performance. This literature study creates an overview of used definitions of an AHE and performance metrics. The study is performed in workup to external validation protocol design for to evaluate the Hypotension Prediction Index (HPI) (Edwards Lifesciences, Irvine, USA) on a data set of intensive care patients. Therefore, the goal is to converge the medical and technical domains by asking the following questions:

- What is the definition of hypotensive event used in development and validation of hypotension prediction algorithms?
- What performance metrics have been used to quantify algorithm performance in hypotension prediction algorithms?

These questions will be answered in the next section. Chapter 2.1 discusses background information that lay the foundation of upcoming chapters. Both hypotension in terms of physiology and different definitions, as well as statistical performance measures will be discussed. Chapter 2.2 describes the used methodology in finding relevant literature. Chapter 2.3 presents definitions of hypotension used in predictive models. Chapter 2.4 presents the results on used performance metrics for discriminative abilities of the model, i.e., how well it can separate the events from the non-events. Chapter 2.5 outlines measures used in previous studies to quantify calibration of predictive models, i.e., the relation between predicted probability and actual occurrence of events. Chapter 2.6 discusses measures used to express the timeliness of alarms, i.e., model performance for different temporal alarm lead times. Lastly, Chapter 2.7 forms the discussion and defines the knowledge gap for the master thesis that follows this review.

2

Literature Study

This chapter summarises literature to answer the previously stated research questions. First, background information is provided that form the theoretical foundation of later discussed concepts. In the sections thereafter

2.1. Background Theory

2.1.1. Blood pressure and Hypotension

To understand the possible tissue damage resulting from hypotension, the concept of oxygen and nutrition delivery to tissue must be understood. Oxygen and nutrition delivery to tissue depend on multiple complex factors in critically ill patients. However, a general determinant of delivery is blood flow, which can be approximated by arterial blood pressure divided by vascular resistance.

$$\text{Blood Flow} = \frac{\text{Blood Pressure}}{\text{Vascular Resistance}} \tag{2.1}$$

This shows that blood pressure drives blood flow. Local blood flow to organs is autoregulated by adjusting the local vascular resistance. By autonomous adaptation of muscle tone in special vascular beds, a constant blood flow is maintained upon changes in blood pressure. However, blood flow regulation can only be achieved within the autoregulatory range of blood pressures for a particular organ. Decrease of blood pressure below the autoregulatory threshold leads to inadequate blood flow and thus insufficient oxygen delivery, to which the heart, brain and kidneys are particularly vulnerable to.[16]

On a cellular level, hypoxia is damaging in multiple ways. Energy metabolism in mitochondria is switched from aerobic to anaerobic glycolysis. This procures an excess amount of lactate, hydrogen ions and inorganic phosphates. As adenine triphosphate (ATP) levels decrease because of diminished production and continued consumption, protein synthesis is hampered. This imperils mitochondrial function and will eventually lead to activation of apoptosis and end-organ failure.[17,18]
Cell damage can occur when blood pressure drops below the autoregulatory threshold of tissue. Inadequate local blood flow leads to hypoxia, which induces anaerobic metabolism, mitochondrial failure and end-organ failure.[18] Physiological thresholds of hypotension differ per patient and per organ.

Definition of Hypotension (and Hypotensive events)

No consensus is reached on the clinical definition of hypotension.[19] Many different definitions for hypotension exists, with criteria using absolute systolic (SBP), diastolic (DBP) or mean arterial blood pressure (MAP) values and their relative differences to the baseline values of the patient.[19] A commonly used threshold of hypotension is a mean arterial blood pressure of 65 mmHg.[19] This originates form experiments in mammals that indicated a lower cerebral autoregulatory threshold at 65 mmHg.[20,21] However, autoregulatory thresholds differ per organ and per patient.[22,23] For example, patients with a history of chronic hypertension can present with shock symptoms under normotensive blood pressure.[2] So, no perfect absolute threshold exists.

2.1.2. Validation

In addition to clinical background information, it is also important to elaborate on statistical concepts that form the foundation of the literature study. Quality of a predictive model needs to be determined after development. This process, called validation, comes down to assessing the model on generalizability, i.e., the agreement between predictions of the model and actual outcome in reality. Conventional validation of a model can be divided into two different quality aspects: discriminative performance and calibration of the model. Both aspects can be described by different performance metrics. Also 'overall' performance metrics exist that attempt to comprise both aspects into a single numerical value.

Theory behind Discrimination

Discriminative ability refers to the ability of a predictive model to separate those that will develop a certain outcome from those that will not. Therefore, all data points need to be labelled on predicted outcome and actual outcome. The binary labels are illustrated in Table 1. True Positives (TP) are correctly predicted events. True Negatives (TN) are correctly predicted non-events. False Positives (FP) are falsely predicted non-events, also known as type I errors. False Negatives (FN) are events that were missed, are also known as type II errors. These classifications can be listed into a table called the confusion matrix or contingency table, as shown in Table 1. This is a useful tool for model performance assessment that forms the foundation of other performance metrics.[24] Various methods exist to quantify discriminative performance.

Sensitivity and Specificity Two of the most popular performance metrics for binary classes in the clinical domain are sensitivity and specificity. Sensitivity, also commonly referred to as Recall of positive class, is the true positive rate or the share of TP of all events,TP/(TP+FN). This metric contains the information on the chance of missing alarms for true events. Specificity is defined as the true negative rate or the share of TN of all non-events, TN/(TN+FP). This measure contains information on the chance that the alarm is a correct prediction.

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.3}$$

Positive and negative predictive values Positive Predictive Value (PPV), also commonly referred to as Precision, is the share of TP in all alarms, TP/(TP+FP). It thus provides the important information on the probability that the alarm will actually be followed by an event.[15] PPV can be regarded as a measure for how exact the positive predictions are.[25] Negative Predictive Value (NPV) is the share of TN in all non-alarms, TN/(TN+FN). It explains the probability that a non-alarm is followed by a non-event.

$$PPV = Precision = \frac{TP}{TP + FP} \tag{2.4}$$

$$NPV = \frac{TN}{TN + FN} \tag{2.5}$$

The benefit is the metrics above is that is provides an incredibly useful tool to report performance of tests with binary outcomes and of its wide use forms the foundation of statistical analyses.[15]

Accuracy This widely used statistical metric is often used as a conclusive measure of discrimination. Accuracy is the proportion of correct predictions in all predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{\#\, correct\ predictions}{\#\, samples} \tag{2.6}$$

Note that TPs have the same weight as TNs in the calculation of accuracy. In an imbalanced data set, i.e., having a skewed classification distribution, accuracy becomes biased.[26] This can be illustrated by the following example: For a data set with a class distribution of 95% non-events, a naively predictive model can predict every sample as a non-event to obtain an accuracy of 95% percent.[25] This level of accuracy appears to be impressive, whilst none of the actual events is predicted correctly.

Table 2.1: Confusion matrix by Kubben *et al.*[24]. Abbreviations: TPR, true positive rate; TNR, true negative rate.

		Observation			
		True	False		
Prediction	True	True positive (TP)	False positive (FP)	→	Positive predictive value (PPV)
	False	False negative (FN)	True negative (TN)	→	Negative predictive value (NPV)
		↓ Sensitivity (TPR)	↓ Specificity (TNR)		

Receiver Operating Characteristic Curve A frequently used method for analysing discriminative performance is the receiver operating characteristic (ROC) curve. An ROC curve displays the performance of the model by plotting the sensitivity against 1-specificity for all threshold values.[27] A point on the ROC curve represents model performance for a single, particular threshold value. Thereby, true positive rate is plotted against false positive rate for the complete range of thresholds. This gives plot gives an overview of model performance. Thereafter, the ROC curve can also be used to pick the eventual classification threshold, in combination with a cost function. A cost function defines the desired trade-off between sensitivity and specificity, which also forms a line on the ROC plot. The intersection of the ROC curve and cost function indicate the optimal threshold value for that cost function. But also without any cost function, the ROC curve shows valuable information when the operating threshold of the algorithm is yet to be determined.[26] Via the ROC curve, discriminative ability of a model can be summarised in a single numerical measure: the area under the curve (AUC). The area under the ROC curve (AUROC) is a common technique for evaluating and comparing models on datasets with varying class distribution.[26] A perfect model has an AUC of 1. A non-informative model has an AUC of 0.5, which is the case when predicting 'heads' of 'tails' when flipping a coin.

Precision-Recall Curve Precision, also known as PPV, entails the exactness of positive predictions. It is acquired by dividing the number of TPs by the total number of positive predictions. Recall of the positive class, also known as sensitivity, entails the completeness of which events were correctly predicted. It is acquired by dividing the number of TPs by the total number of events. The relation between Precision and Recall can be illustrated by plotting the two against each other for the complete range of classification thresholds. This is analogous to the ROC curve.[26] Recall and Precision are similar to accuracy and error, respectively. But both accuracy and error are sensitive to changes in data distributions, i.e., the ratio between events and non-events. Regarding precision and recall, only precision is subjective to the data distribution and recall is not. Therefore algorithm performance can be effectively evaluated by using precision and recall when applied to unbalanced data sets.[25] The area under the PR curve (PR-AUC) allows comparison of multiple models using a single numeric value, similar to the AUROC.

F-Measure The F-measure is an alternative on accuracy and displays classification performance in terms of a weighed ratio between recall and precision. It is most used with factor beta equalling 1, under the name 'F1-score'.

$$F - measure = \frac{(1+\beta)^2 \times Recall \times Precision}{\beta^2 \times Recall + Precision} \tag{2.7}$$

Theory behind Calibration

Calibration is the agreement between the predictions and observed outcomes and is one of the primary requirements to determine clinical usefulness.[15] Perfect calibration is reached when for every value of

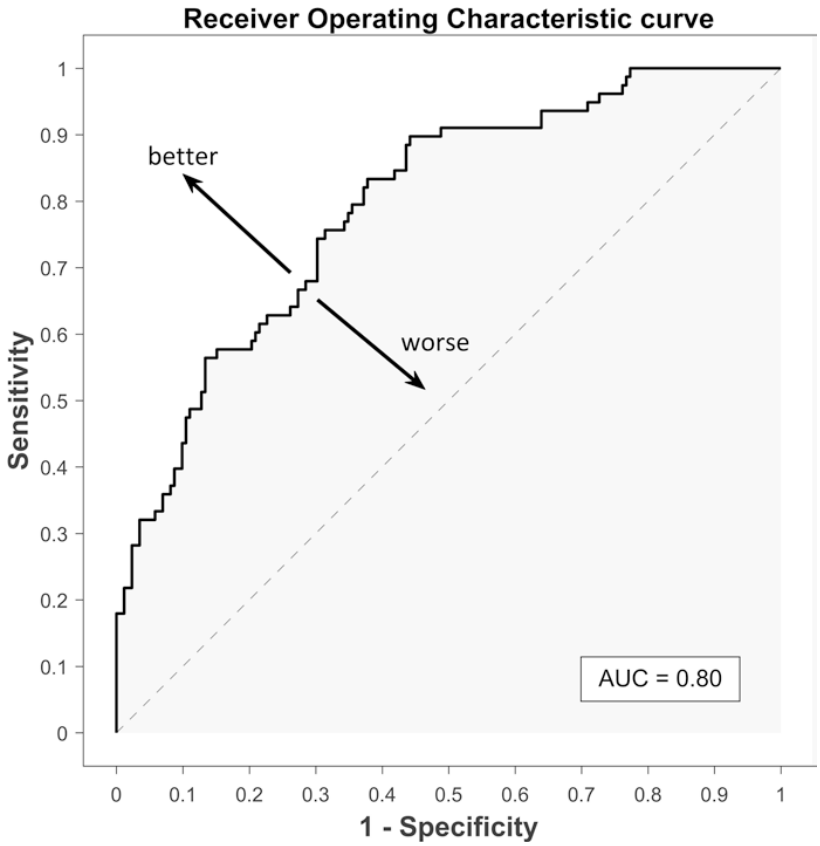


Figure 2.1: *ROC curve by Kubben* et al. [24]*. A point on the solid line represents the values of sensitivity and specificity for one single threshold value. An uninformative model assigns random predictions in the same ratio of classes as observed in the underlying population. In that case, sensitivity equals 1-specificity, the ROC curve would follow the dashed line and AUC would be 0.5. Therefore, to perform better than chance, the ROC-curve of the model needs to be located above the dashed line. Abbreviations: AUC, area under the curve. Figure subject to Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/)*

calculated probability, the actual event occurs in the same ratio. For example, in the group of patients for which the model estimated the probability on event was 80%, the actual occurrence of event was also 80%.

Hosmer-Lemeshow statistic To evaluate calibration according to the Hosmer-Lemeshow statistic, the complete range of predicted probability values is divided into subgroups equally. Per subgroup, the absolute number and proportion of observed events is displayed. The Hosmer-Lemeshow statistic captures the overall goodness-of-fit into a single numerical number and provides p-values for subgroups on the null-hypothesis of a good fit. However, this statistic has a low power and does not show the direction of miscalibration.[28]

Calibration curve Calibration quality can also be displayed visually, as illustrated in 2.2. Similar to the Hosmer-Lemeshow test, groups are divided on predicted probability range. But now, average actual occurrence of events - optionally with error bars - are plotted over probability values. Two parameters can be extracted that describe the level of calibration: calibration slope and calibration-in-the-large, which is the intercept of the curve with the y-axis for a slope set to 1.[15,24]

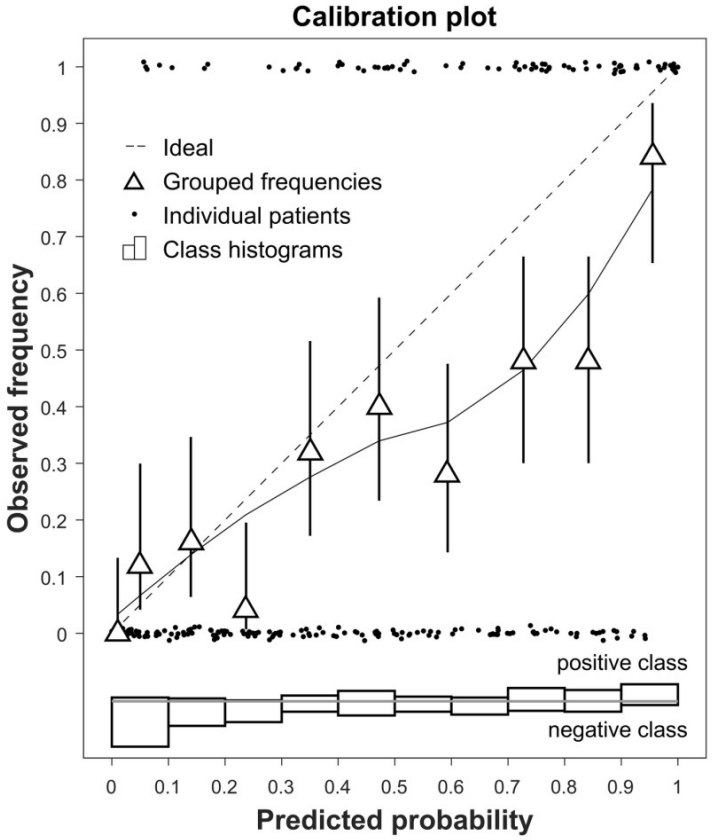


Figure 2.2: *Calibration curve by Kubben* et al.[24]*. For a perfectly calibrated model, the distribution of the outcomes for a value of predicted probability exactly matches this value. Hence, the curve follows the dashed line. Black dots represent outcome in values of 1 or 0, for events and non-events, respectively. Figure subject to Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/)*

Theory behind overall performance measures

'Overall performance metrics' capture quality of discriminative ability, as well as calibration into a single numerical value. Overall performance metrics define general measure of "goodness of fit". This comes conceptually down on the 'distance', or difference, between predicted and observed outcomes.[15] For example, in weather forecasting the distance is the difference between predicted and measured maximum temperature that day.[15] When model output is the probability on a binary outcome, as in event vs. non-event, distance is defined as the calculated probability of event occurrence p minus the outcome (1 for event, 0 for non-event). For example, if a model predicts a 90% chance of event occurrence correctly, the distance is 0.90 – 1 = 0.10. The smaller the distance, the better the fit of the predictive model.[15]

Brier Score The Brier-score in equation 2.8 is the mean squared error of the forecast (f), or probability of event, in which o represents the outcome. A lower Brier-score represents a better goodness-of-fit. However, this score is sensitive to the occurrence rate of the event. This is illustrated by the following scenario: A non-informative, and thus useless, model simply could return a probability that is equal to the rate of occurrence of the event. If an event occurs for 50% of the data points, the non-informative model will always give a probability of 50% for occurrence of the event. With an event rate of 50%, the maximum, worst, Brier-score score is $(0.50–0)^2 = 0.25$. But, for an event rate of 10%, the maximum score is $(0.10–0)^2 = 0.01$. Therefore, the Brier-score is susceptible to bias due to class imbalance in

the data set. This could be partially corrected by scaling of the Brier-score with the maximum non-informative score given the class balance of a particular data set.

$$Brier - score = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{2.8}$$

Internal vs. external validation
Internal validation is often a first step after model development to test the model on reproducibility and is an indication for generalizability. Internal validation uses data that was kept apart from the originally available data. This data is exclusively used for testing and is not used in model development to avoid biased test results. As internal validation data originates from the same underlying population as the training data, it is sensitive to confounding. Therefore, internal validation leads to a minimal indication of model generalizability, as it only allows conclusions on the directly underlying population the model is trained on.[15]

Generalizability can only be assessed via external validation, by using an independent data set.[29] Data for external validation can differ in geographic location (e.g., hospital or country), moment in time and clinical case type or severity.[28] Thereby, the model is placed in a broader context, which shows true model performance that can be expected upon implementation.

2.2. Literature search method

The search methodology started 'organically' by using papers reporting on HPI as a start point and manually searching reference lists for other relevant articles. However, a more systematic approach was favoured for this literature study to find the articles on validation of hypotension algorithms. A search was performed in the PubmMed database on March 23 2021 using the query *("Hypotension"[MeSH] AND "Algorithms"[MeSH]) OR (hyptension AND algorithms AND machine learning)*. Additionally, backward and forward snowballing was applied, according to Wohlin [30], by manually searching through reference lists and citations of all included articles. This could also include articles from other databases, such as IEEE. Relevant articles were screened on title, abstract or full text on development or validation of hypotension prediction models. Reports on hypotension prediction models because of major interventions were excluded (e.g., induction of anaesthesia or haemodialysis).

Initially, 246 results were yielded by the search query, of which 17 articles were included. Both forward and backward snowballing yielded another 11 articles in three iterations. A total of 28 articles were included in the literature study. All studies are detailed in Table 2.2 and Table 2.3.

2.3. Hypotension definition

In validation of hypotension prediction algorithms, various definitions of hypotension for ICU patients are used. A complete overview is provided in Table 2.2. The first of the two most commonly used definitions was stated by Moody & Lehman [31] in the PhysioNet/Computers in Cardiology Challenge 2009. This challenge for development of models to predict acute hypotensive events (AHEs) defines an AHE as a period of mean arterial pressure (MAP) < 65 mmHg for ≥ 90% of the time within a 30-minute window. This definition was not only adopted by participants of the challenge, but also by other developers of predictive models to explicitly allow comparison with other models described in literature.[32–38] A total of 19 studies (68%) used the definition of the PhysioNet Challenge for an AHE. Additionally, Moghadam *et al.*[32] defined non-hypotension as data points with MAP < 75 mmHg that occurred ≥ 40 min before or ≥ 20 minutes after a hypotensive event.

The second most commonly used definition of hypotension is used by the commercially available Hypotension Prediction Index (HPI) by Hatib *et al.*[9]. The HPI-algorithm defines hypotension as a period of MAP < 65 mmHg for ≥ 1 minute. Non-hypotension is defined as MAP > 75 mmHg. Data segments were excluded if MAP decreased with more than 0.5 mmHg/s as this would result from unpredictable external manipulation (e.g., blood loss or pressure transducer repositioning).[9] External validation of the HPI on noncardiac surgical patients by Davies *et al.*[39] copied this definition and also excluded

data points of suspected haemodynamic intervention to prevent false negative bias. Segments with an increase of MAP >5 mmHg in 20 seconds or > 8 mmHg in 2 minutes were excluded from analysis. Validation of HPI during cardiac surgery by Shin *et al.* [40] also used the definition above for hypotension, but defined non-hypotension as MAP > 65 mmHg for > 1 minute. Shin *et al.* [40] used an additional definition of a severe hypotensive event as MAP < 50 mmHg for > 1 minute.

Similarly to the above definition, Cherifa *et al.* [41] used MAP < 65 mmHg for > 5 minutes as definition of hypotension in development of their predictive model. Other definitions included two blood pressure conditions in the definition of hypotension. Donald *et al.* [10] used systolic blood pressure (SBP) < 90 mmHg or MAP < 70 mmHg for > 5 minutes in a model on traumatic brain injury induced hypotension. Yoon *et al.* [42] defined hypotension as both SBP < 90 mmHg and MAP < 60 mmHg for more than 50% of the time within a 10 minute window.

A patient specific definition of hypotension was only used by Chan *et al.* [43], that used relative thresholds. Here, a difference between the 5-minute and 60-minute moving average of the MAP was used to define the onset of hypotension. A drop of the 5-minute moving average below 20% of the 60-minute moving average of MAP was used to define the onset of a hypotensive event.

The variety in definitions of hypotensive events used for predictive algorithms comes from the lack of consensus on hypotension in medical literature. As an example, Bijker *et al.* [19] concluded from a systematic literature study that the variety of intraoperative hypotension (IOH) differs dramatically. Applying the most extreme definitions of hypotension onto a cohort of patients undergoing noncardiac surgery led to incidences of IOH between 5% and 99%. It was stated that a workable and proper definition of IOH should include: threshold value and type (absolute or relative), baseline (for relative), blood pressure type (systolic vs. mean), measurement method and interval, and minimal episode duration. IOH was stressed to be a dynamic phenomenon of which the threshold and event duration could be defined according to the prevalence of different patient outcomes (e.g., myocardial ischaemia, ischemic stroke, “watershed” infarction). But on the contrary, they suggested that hypotension should not be defined using a static threshold, as all dynamic factors of hypotension should be considered in its definition.

The Society of Critical Care Medicine’s Surviving Sepsis Guidelines from 2016 by Rhodes *et al.* [44] recommends a MAP of 65 mmHg as target for initial resuscitation for septic patients requiring vasopressor use. Whereas recommendations from 2012 explicitly state a target mean arterial blood pressure of ≥ 65 mmHg.[45]

Risk on complications increases with depth and duration of hypotension, as concluded by a systematic review by Wesselink *et al.* [46], on the risk of postoperative outcomes upon IOH. Relative risk of end-organ damage in noncardiac surgery started to increase for a mean arterial hypotension < 80 mmHg for an exposure of ≥ 10 minutes. Associations between intraoperative hypotension and postoperative complications were also observed in a meta-analysis by Wijnberge *et al.* [47]. IOH led to on odds-ratio (with 95% confidence interval) on cardiac outcomes of 2.44 (1.52 to 3.93), on acute kidney injury of 2.69 (1.31 to 5.55) and on mortality of 1.94 (1.32 to 2.84).[47]

2.4. Discrimination

The use of discrimination metrics used in papers reporting on predictive models for AHEs is presented in Table 2.3. A large variety of metrics is used to express discriminative performance. Performance metrics we limited to sensitivity, specificity and accuracy in 9 of the 28 (32%) of the included studies. Rocha *et al.* [48] supplemented these three metrics with the correlation coefficient to evaluate the prediction of continuous MAP values. Two other studies also used a limited number of metrics to express discriminative performance. The F1-score as a performance metric for the developed predictive model was the only used metric by Pathinarupothi & Rangan [49]. Sensitivity, positive predictive value (PPV) and the normalised root mean square of error (NRMSE) were the only metrics used by Ghaffari & Jalali [50]. NRMSE is de facto a the sqaure root of the Brier-score.

A large arsenal of metrics was used by Cherifa *et al.* [41], who validated their predictive machine learning algorithm by reporting the Brier score, are under the receiver operator characteristic curve (AUROC), sensitivity, specificity, PPV, negative predictive value (NPV), and positive and negative likelihood ratios. The receiver operator characteristic curve (ROC) was used in 8 of the 28 studies (29%) and AUROC was used in 7 studies (25%). Surprisingly, these two metrics were used mutually exclusively in all but one of the mentioned studies.

All validation studies on HPI reported ROC curves in addition to the sensitivity, specificity and PPV.[9,39,51,52] Of these, only the external validation by Ranucci *et al.* [51] included lines of 95% confidence interval to the ROC curve. Similarly, ROC curves with error bars were reported Donald *et al.* [10] on a different model. Other discriminative performance metrics included the F1-score, the precision-recall curve, and the AUC of the PR-curve.

2.5. Calibration

Only 5 studies (18%) reported on calibration of hypotension prediction algorithm specifically. A calibration curve was used in 4 studies (14%). A table was used twice to display the distributions of time to event against the hypotension probability of the model. None of the articles reported values of calibration slope nor calibration-in-the-large. However, Ranucci *et al.* [51] did estimate the correlation between predicted and observed events with a logistic function, specifically for hypotension prediction index (HPI) values during 5 and 7 minutes prior to the hypotensive event.

2.6. Timeliness

The theoretical benefit of a predictive model increases, if the model is able to alarm for impending hypotension more time in advance. Therefore, timeliness of a predictive model is an important aspect in evaluating clinical usefulness. However, it is not a conventional performance aspect in validation of algorithms, as discrimination and calibration are. Timeliness as a performance aspect should also be considered when estimating clinical usefulness of the model. .

Timeliness of the predictive model was evaluated in 17 articles (61%). An overview of used methods is detailed in Table 2.2. None of the included articles used a form of scoring nor a calculated performance metric that was dedicated to quantification of timeliness, as seen for discrimination of calibration. However, two types of assessment of temporal performance were observed across included articles, here called ‘backward’ and ‘forward’ timeliness assessment.

In backward timeliness assessment, performance metrics for discriminative performance and calibration are displayed for several alarm lead times. In the definition of a TP, alarm lead time is the maximum allowed duration between alarm and onset of hypotension. For example, if the alarm lead time is defined as 15 minutes and an AHE starts 16 minutes after the initial alarm, this alarm is deemed as a false positive. Here, the onset of hypotension occurred too late. By presenting performance metrics for different alarm lead times, model quality under different temporal horizons is displayed.

Backward timeliness assessment was performed in 15 articles (54%). The values for alarm lead time and the accompanied performance metrics depended on the scope of the predictive model. Multiple ROC curves for different alarm lead times were only included in the reports of Davies *et al.* [39] and Shin *et al.* [40].

In forward timeliness assessment, the distribution of duration between alarm and onset of hypotension is displayed. This duration is also known as the time-to-event. Assessment via time-to-event is observational and independent of the earlier defined alarm lead time. The distribution of time-to-event can be presented via a histogram or by providing the values for a number of percentiles. To gain deeper insight in temporal model quality, time-to-event distributions are presented for multiple alarm threshold ranges.

Forward timeliness assessment was performed in 3 articles (11%). Time-to-event values of the

25th, 50th and 75th percentile were provided by Davies *et al.* [39] and Maheshwari *et al.* [52]. A box-plot of time-to-event distribution on the y-axis over alarm threshold values on the x-axis was used by Donald *et al.* [10]. No histograms were used to present time-to-event distributions.

Table 2.2: Used definitions of Hypotension and used metrics of alarm timeliness in hypotension prediction models.

Authors	Year	Hypotension definition	Timeliness assessment
Bhattacharya *et al.* [53]	2014	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (SE, SP and accuracy on 1-120 min)
Bhattacharya *et al.* [36]	2018	MAP <60 mmHg for >90% of 30 min	N/A
Chan *et al.* [43]	2020	MAP drop >20% between 60 min and 5 min moving averages	Metrics on different alarm lead times (ROC on 10, 20 and 30 min) & Heat map of: AUROC vs different alarm lead times vs observation window
Cherifa *et al.* [41]	2020	MAP <65 mmHg for >5 min	N/A
Davies *et al.* [39]	2020	MAP <65 mmHg for >1 min	TTE distribution percentiles (25th ,50th and 75th) for 10 HPI ranges
Dernoncourt *et al.* [38]	2015	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (AUROC on 10-60 min)
Donald *et al.* [10]	2019	SBP <90 or MAP <70 for >5 min	TTE boxplot vs threshold ranges
Ghaffari & Jalali [50]	2009	MAP <60 mmHg for >90% of 30 min	N/A
Ghosh *et al.* [54]	2016	MAP <60 mmHg for >90% of 30 min	N/A
Hatib *et al.* [9]	2018	MAP <65 mmHg for >1 min	Metrics on different alarm lead times (SE and SP on 5, 10 and 15 minutes)
Jiang *et al.* [55]	2017	MAP <60 mmHg for >90% of 30 min	N/A
Jiang *et al.* [56]	2020	MAP <60 mmHg for >90% of 30 min	N/A
Lee & Mark [57]	2010	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (all used on 1-2 h)
Maheshwari *et al.* [52]	2020	MAP <65 mmHg for >1 min	Metrics on different alarm lead times (AUROC, SE and SP on 5, 10 and 15 min) & TTE distr percentiles (25,50,75) for HPI ranges
Moghadam *et al.* [32]	2020	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (no. of TP on 5, 10, 15 and 20 min)
Moghadam *et al.* [35]	2020	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (SE, PPV and F1-score on 1-30 min)
Moghadam *et al.* [58]	2021	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (SE on 1-30 min)
Pathinarupothi & Rangan [49]	2017	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (F1-score on 15-165min)
Pathinarupothi *et al.* [11]	2018	MAP <60 mmHg for >90% of 30 min	N/A
Ranucci *et al.* [51]	2019	MAP <65 mmHg for >1 min	N/A
Rocha *et al.* [59]	2010	MAP <60 mmHg for >90% of 30 min	N/A
Rocha *et al.* [48]	2011	MAP <60 mmHg for >90% of 30 min	N/A
Shin *et al.* [40]	2020	MAP <65 mmHg for >1 min	Metrics on different alarm lead times (AUROC, SE and SP on 5, 10 and 15 min & ROC on 5 and 10 min)
Sun *et al.* [60]	2021	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (RMS on 1-60 min)
Tsuji *et al.* [33]	2020	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (all used on 1-7 min)
Xiao *et al.* [37]	2019	MAP <60 mmHg for >90% of 30 min	N/A
Yoon *et al.* [42]	2020	SBP ≤ 90 mmHg and MAP ≤ 60 mmHg for ≥ 50% of 10 min	Metrics on different alarm lead times (PRAUC and AUROC on 1-300 min)
Zhang *et al.* [34]	2020	MAP <60 mmHg for >90% of 30 min	Metrics on different alarm lead times (AUROC on 30-180 min)

Abbreviations: HPI, hypotension prediction index; PRAUC, area under the precision-recall curve; RMS, root mean square; ROC, receiver operator characteristic curve; SE, sensitivity; SP, specificity; TP, true postive; TTE, time-to-event.

Table 2.3: Used metrics to assess discriminative performance and calibration hypotension prediction models.

Authors	Year	Overall	Discrimination								Calibration	Other
		Brier	SE	SP	PPV	NPV	ACC	F1	ROC	AUROC	Calibration Curve	
Bhattacharya *et al.* [53]	2014		x	x			x					
Bhattacharya *et al.* [36]	2018		x	x			x					
Chan *et al.* [43]	2020				x				x			
Cherifa *et al.* [41]	2020	x	x	x	x	x	x		x			LR+/LR-
Davies *et al.* [39]	2020		x	x	x				x			calibration table
Dernoncourt *et al.* [38]	2015		x		x					x		PPV at 0.90 SE
Donald *et al.* [10]	2019		x	x					x			PR-curve
Ghaffari & Jalali [50]	2009		x		x							NMRS
Ghosh *et al.* [54]	2016		x				x					
Hatib *et al.* [9]	2018		x	x	x	x			x	x	x	
Jiang *et al.* [55]	2017		x	x			x					
Jiang *et al.* [56]	2020		x	x			x					
Lee & Mark [57]	2010		x	x	x	x	x			x		
Maheshwari *et al.* [52]	2020		x	x	x				x		x	
Moghadam *et al.* [32]	2020		x	x	x		x	x		x		
Moghadam *et al.* [35]	2020		x	x			x		x			
Moghadam *et al.* [58]	2021		x	x	x		x	x				
Pathinarupothi & Rangan [49]	2017							x				
Pathinarupothi *et al.* [11]	2018		x	x			x					
Ranucci *et al.* [51]	2019		x	x	x	x			x			le Cessie-van Houwelingen test
Rocha *et al.* [59]	2010		x	x			x					
Rocha *et al.* [48]	2011		x	x			x					Correlation coefficient ABP forecast
Shin *et al.* [40]	2020		x	x	x	x				x		
Sun *et al.* [60]	2021		x	x	x							RMS
Tsuji *et al.* [33]	2020		x	x			x			x		
Xiao *et al.* [37]	2019		x	x			x					
Yoon *et al.* [42]	2020	x	x		x					x	x	AUPRC
Zhang *et al.* [34]	2020		x		x		x	x		x		

Abbreviations: ACC, accuracy; AUROC, area under the ROC curve; LR, likelihood ratio; NPV, negative predictive value; NRMS, normalised root mean square; PPV, positive predictive value; PR-curve, precision-recall curve; RMS, root mean square; ROC, receiver operator characteristic; SE, sensitivity; SP, specificity.

2.7. Discussion

Regarding hypotension, a variety of definitions are used in development and validation of hypotension prediction algorithms. This is a result of the lack of consensus on the clinical definition of hypotension. One must be aware that predictive models on hypotension can be developed with slightly different goals or requirement and therefore use different definitions. Generally, two definitions of hypotension were used for hypotension prediction models. The two definitions differ substantially in required sub-threshold duration (> 1 min vs >27 min below threshold). This hampers comparison of the algorithms.

The vast majority of the articles used static definitions of hypotension. This prevents the ability of setting patient-specific blood pressure targets. While absolute thresholds may be more objective and convenient for model training, ability to set a relative or patient-specific threshold remain a potential area of improvement for hypotension prediction models.

In assessment of discriminative performance, a large variety of metrics was used. Apart from sensitivity and specificity, accuracy was the mostly used metric, included in more than half of the articles. However due to both absolute class imbalance in a validation data set and differences in class distribution between data sets, accuracy can be heavily biased. This makes accuracy useless for inter- and intra-model comparison in validation.

Metrics on discriminative performance we often consisting of sensitivity, specificity and (AUC of) ROC curves. These measures are proven to be fairly robust to underlying class imbalance, which enables comparison between algorithms and validation studies. However, use of ROC curves may be encouraged in addition to reporting AUROC alone, as this gives even more valuable information on model quality. Also, may the display of threshold values in the ROC curve be informative for commercially available algorithms. This gives clinical insight in the meaning of the predictive model output that is displayed bed-side.

Evaluation of calibration of algorithms was often overlooked. Despite other studies judging calibration visually via a plot or table, only Ranucci *et al.* [51] quantified calibration. Calibration of a single predictive model can be judged visually.[15] But additional quantification would allow more accurate comparison between predictive models, as well as the identification of subgroups eliciting altered performance.

Performance measures with varying degree in robustness to class imbalance have been used in previous research validating hypotension prediction algorithms. Other specifically informative performance measures exist to assess model quality and clinical implications. More insight in the clinical usefulness of predictive models can be provided by precision-recall curves or extra focus on temporal quality of models. For example, Dernoncourt *et al.* [38] and Solomon *et al.* [61] calculated PPV under alarm thresholds that were set to match certain specificity levels. This provided more insight on occurrence of false predictions. Other examples of robust measures that could be used more often are Matthew's Correlation Coefficient and F-measure.[62]

Several data processing steps in used validation methodology reduce the clinical generalizability of the hypotension prediction algorithm. Some studies exclude data samples from a 'grey' zone of MAP 65-75 mmHg, this reduces the amount of false predictions and heavily reduces the reliability of reported performance.[9,52] Discrimination and distribution of predictions near the alarm threshold are determinant for clinical usefulness.[15]

Overoptimism was noted in studies reporting performance on non-clinical alarm thresholds. Hatib *et al.* [9] and Davies *et al.* [39] calculated an extensive amount of performance measures with the statistically optimal alarm threshold of HPI, of around 40. However, the alarm threshold of the commercially available HPI is set at 85. This substantial difference in threshold begs for a correction of sensitivity and timeliness of the model.

Novel continuous evaluation methods for validation, as described by Moghadam and colleagues[35,58] mimic real-time monitoring performance and are potentially a more realistic method of validation. All other studies apply heavier discretization of the data by segmenting data into multi-minute time windows to label these as TP, FP, etc. This discrete, tumbling window approach uses non-overlapping

time windows. Continuous, sliding window evaluation handles labels each minute or each individual model prediction as TP, FP, etc. The label then depends on time period following this individual minute or prediction. The continuous, sliding window approach mimics a real-life clinical situation on the ICU with real-time monitoring. The continuous evaluation method leads to more labels than the discrete evaluation method for the same amount of patient data. This may lead to relatively more negative predictions. A more imbalanced data set negatively affects the PPV. However, PPV and other performance metrics resulting from continuous validation would give a more realistic representation of real-life performance of the hypotension predicting algorithm.

2.7.1. Knowledge gap

The knowledge gap identified in this literature study regard the lack of consensus on class definition and validation methodology.

- What defines a true positive prediction?
  Class definition is fundamental of model validation. However, previous research has not always been transparent on data processing and discretization of continuous data to enable classification.
- What is a clinically relevant balance between sensitivity, specificity, PPV and timeliness?
  Algorithm design requires picking a threshold value to which the model decides to alarm for hypotension or not. No trade-offs between these aspects have been explicitly published for threshold selection.
- Therefore, what is the cost function to determine a clinically relevant threshold ?
- How can maximize transfer of information on model performance to the reader with a clinical background, by using non-conventional or dedicated performance metric? s to improve judgement on clinical usefulness of the model, but without losing greater public?
  A limited arsenal of statistical measures has been used to describe model quality of hypotension prediction algorithms. Presumably, validation method complexity is curbed to increase readability of the article for the clinical reader. This happens at the expense of information to estimate clinical usefulness on.

These knowledge gaps forms the basis of the upcoming master thesis. Planning and proposed methodology on the master thesis project will follow in the master thesis work plan.

References

1. Bagshaw, S. M. *et al.* A Multi-Center Evaluation of Early Acute Kidney Injury in Critically Ill Trauma Patients. *Renal Failure* **30,** 581–589. doi:10.1080/08860220802134649 (Jan. 2008).
2. Gamper, G. *et al.* Vasopressors for hypotensive shock. *Cochrane Database of Systematic Reviews* **2016.** doi:10.1002/14651858.CD003709.pub4 (Feb. 2016).
3. Maheshwari, K. *et al.* The relationship between ICU hypotension and in-hospital mortality and morbidity in septic patients. *Intensive Care Medicine* **44,** 857–867. doi:10.1007/s00134-018-5218-5 (June 2018).
4. Luo, W. *et al.* Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research* **18.** doi:10.2196/jmir.5870 (Dec. 2016).
5. Sun, L. Y. *et al.* Association of intraoperative hypotension with acute kidney injury after elective noncardiac surgery. *Anesthesiology* **123,** 515–523. doi:10.1097/ALN.0000000000000765 (Sept. 2015).
6. Seymour, C. W. *et al.* Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (sepsis-3). *Journal of the American Medical Association* **315,** 762–774. doi:10.1001/jama.2016.0288 (Feb. 2016).
7. Leone, M. *et al.* Optimizing mean arterial pressure in septic shock: A critical reappraisal of the literature. *Critical Care* **19.** doi:10.1186/s13054-015-0794-z (Dec. 2015).
8. Saugel, B. *et al.* Predicting hypotension in perioperative and intensive care medicine. *Best Practice and Research: Clinical Anaesthesiology* **33,** 189–197. doi:10.1016/j.bpa.2019.04.001 (June 2019).
9. Hatib, F. *et al.* Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. *Anesthesiology* **129,** 663–674. doi:10.1097/ALN.0000000000002300 (Oct. 2018).
10. Donald, R. *et al.* Forewarning of hypotensive events using a Bayesian artificial neural network in neurocritical care. *Journal of Clinical Monitoring and Computing* **33,** 39–51. doi:10.1007/s10877-018-0139-y (Feb. 2019).
11. Pathinarupothi, R. K. *et al. Deriving High Performance Alerts from Reduced Sensor Data for Timely Intervention in Acute Hypotensive Episodes* in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* **2018-July** (Institute of Electrical and Electronics Engineers Inc., Oct. 2018), 3260–3263. doi:10.1109/EMBC.2018.8512945.
12. Eshelman, L. J. *et al.* Development and evaluation of predictive alerts for hemodynamic instability in ICU patients. *Annual Symposium proceedings / AMIA Symposium* **2008,** 379–383 (2008).
13. Kendale, S. *et al.* Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. *Anesthesiology* **129,** 675–688. doi:10.1097/ALN.0000000000002374 (2018).
14. Wijnberge, M. *et al.* Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension during Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. *Journal of the American Medical Association* **323,** 1052–1060. doi:10.1001/jama.2020.0592 (Mar. 2020).
15. Steyerberg, E. *Clinical Prediction Models* doi:10.1007/978-0-387-77244-8 (Springer New York, New York, NY, 2009).
16. Boron, W. F. & Boulpaep, E. L. *Medical Physiology - A Cellular and Molecular Approach* 2nd ed. (Saunders Elsevier, 2012).

17. Antonelli, M. *et al.* Hemodynamic monitoring in shock and implications for management: International Consensus Conference, Paris, France, 27–28 April 2006. *Intensive Care Medicine* **33,** 575–590. doi:10.1007/s00134-007-0531-4 (Apr. 2007).
18. Harten, J. & Kinsella, J. Perioperative optimisation. *Scottish Medical Journal* **49,** 6–9. doi:10.1177/003693300404900102 (Feb. 2004).
19. Bijker, J. B. *et al.* Incidence of Intraoperative Hypotension as a Function of the Chosen Definition Literature Definitions Applied to a Retrospective Cohort Using Automated Data Collection. *Anesthesiology* **107,** 213–220. doi:https://doi.org/10.1097/01.anes.0000270724.40897.8e (2007).
20. Finnerty, F. A. *et al.* Cerebral hemodynamics during cerebral ischemia induced by acute hypotension. *The Journal of clinical investigation* **33,** 1227–1232. doi:10.1172/JCI102997 (Sept. 1954).
21. Fitch, W. *et al.* Effects of decreasing arterial blood pressure on cerebral blood flow in the baboon. Influence of the sympathetic nervous system. *Circulation Research* **37,** 550–557. doi:10.1161/01.RES.37.5.550 (1975).
22. Strandgaard, S. Autoregulation of cerebral blood flow in hypertensive patients. The modifying influence of prolonged antihypertensive treatment on the tolerance to acute, drug induced hypotension. *Circulation* **53,** 720–727. doi:10.1161/01.CIR.53.4.720 (1976).
23. Kato, R. & Pinsky, M. R. Personalizing blood pressure management in septic shock. *Annals of Intensive Care* **5,** 41. doi:10.1186/s13613-015-0085-5 (2015).
24. Kubben, P. *et al. Fundamentals of Clinical Data Science* 1–219. doi:10.1007/978-3-319-99713-1 (Springer International Publishing, Dec. 2018).
25. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21,** 1263–1284. doi:10.1109/TKDE.2008.239 (Sept. 2009).
26. He, H. & Ma, Y. *Imbalanced Learning* (eds He, H. & Ma, Y.) 1–210. doi:10.1002/9781118646106 (Wiley, June 2013).
27. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30,** 1145–1159. doi:10.1016/S0031-3203(96)00142-2 (July 1997).
28. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *European Heart Journal* **35,** 1925–1931. doi:10.1093/eurheartj/ehu207 (Aug. 2014).
29. Justice, A. C. *et al.* Assessing the generalizability of prognostic information. *Annals of Internal Medicine* **130,** 515–524. doi:10.7326/0003-4819-130-6-199903160-00016 (Mar. 1999).
30. Wohlin, C. *Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering* in *EASE '14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (2014). doi:10.1145/2601248.2601268.
31. Moody, G. B. & Lehman, L. H. Predicting acute hypotensive episodes: The 10th annual PhysioNet/Computers in Cardiology Challenge. *Computers in Cardiology* **36,** 541–544 (2009).
32. Moghadam, M. C. *et al. Supervised Machine-Learning Algorithms in Real-time Prediction of Hypotensive Events* in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* **2020-July** (Institute of Electrical and Electronics Engineers Inc., July 2020), 5468–5471. doi:10.1109/EMBC44109.2020.9175451.
33. Tsuji, T. *et al.* Recurrent probabilistic neural network-based short-term prediction for acute hypotension and ventricular fibrillation. *Scientific Reports* **10.** doi:10.1038/s41598-020-68627-6 (Dec. 2020).
34. Zhang, G. *et al.* A machine learning method for acute hypotensive episodes prediction using only non-invasive parameters. eng. *Computer Methods and Programs in Biomedicine* **200,** 105845. doi:10.1016/j.cmpb.2020.105845 (Mar. 2020).
35. Moghadam, M. C. *et al.* A machine-learning approach to predicting hypotensive events in ICU settings. eng. *Computers in Biology and Medicine* **118,** 103626. doi:10.1016/j.compbiomed.2020.103626 (Mar. 2020).

36. Bhattacharya, S. *et al.* A dual boundary classifier for predicting acute hypotensive episodes in critical care. *PLOS ONE* **13** (ed Luo, F.) e0193259. doi:10.1371/journal.pone.0193259 (Feb. 2018).

37. Xiao, G. *et al.* AHE Detection with a hybrid intelligence model in smart healthcare. *IEEE Access* **7,** 37360–37370. doi:10.1109/ACCESS.2019.2905303 (2019).

38. Dernoncourt, F. *et al. Gaussian process-based feature selection for wavelet parameters: Predicting acute hypotensive episodes from physiological signals* in *Proceedings - IEEE Symposium on Computer-Based Medical Systems* **2015-July** (Institute of Electrical and Electronics Engineers Inc., July 2015), 145–150. doi:10.1109/CBMS.2015.88.

39. Davies, S. J. *et al.* Ability of an Arterial Waveform Analysis–Derived Hypotension Prediction Index to Predict Future Hypotensive Events in Surgical Patients. *Anesthesia & Analgesia* **130,** 352–359. doi:10.1213/ANE.0000000000004121 (Feb. 2020).

40. Shin, B. *et al.* Utility of the Hypotension Prediction Index During Cardiac Surgery. *Journal of Cardiothoracic and Vascular Anesthesia*. doi:10.1053/j.jvca.2020.12.025 (Dec. 2020).

41. Cherifa, M. *et al.* Prediction of an acute hypotensive episode during an ICU hospitalization with a super learner machine-learning algorithm. *Anesthesia and Analgesia* **130,** 1157–1166. doi:10.1213/ANE.0000000000004539 (2020).

42. Yoon, J. H. *et al.* Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit. *Critical Care* **24,** 661. doi:10.1186/s13054-020-03379-3 (Dec. 2020).

43. Chan, B. *et al.* Generalizable deep temporal models for predicting episodes of sudden hypotension in critically ill patients: a personalized approach. *Scientific Reports* **10,** 11480. doi:10.1038/s41598-020-67952-0 (2020).

44. Rhodes, A. *et al.* Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Critical Care Medicine* **45,** 486–552. doi:10.1097/CCM.0000000000002255 (Mar. 2017).

45. Dellinger, R. P. *et al.* The Surviving Sepsis Campaign Guidelines Committee including The Pediatric Subgroup* Surviving Sepsis Campaign: International Guidelines for Management of Severe Sepsis and Septic Shock, 2012. *Intensive Care Medicine* **39,** 165–228. doi:10.1007/s00134-012-2769-8 (2013).

46. Wesselink, E. M. *et al.* Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. *British Journal of Anaesthesia* **121,** 706–721. doi:10.1016/j.bja.2018.04.036 (Oct. 2018).

47. Wijnberge, M *et al.* Association of intraoperative hypotension with postoperative morbidity and mortality: systematic review and meta-analysis. *BJS open* **5.** doi:10.1093/bjsopen/zraa018 (2021).

48. Rocha, T. *et al.* Prediction of acute hypotensive episodes by means of neural network multi-models. *Computers in Biology and Medicine* **41,** 881–890. doi:10.1016/j.compbiomed.2011.07.006 (Oct. 2011).

49. Pathinarupothi, R. K. & Rangan, E. S. *Consensus motifs as adaptive and efficient predictors for acute hypotensive episodes* in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* **2017** (Institute of Electrical and Electronics Engineers Inc., Sept. 2017), 1688–1691. doi:10.1109/EMBC.2017.8037166.

50. Ghaffari, A. & Jalali, A. Predicting acute hypotensive episodes based on hr baroreflex model estimation. *Cardiovascular Engineering* **9,** 161–164. doi:10.1007/s10558-009-9087-y (Dec. 2009).

51. Ranucci, M. *et al.* Discrimination and calibration properties of the hypotension probability indicator during cardiac and vascular surgery. *Minerva Anestesiologica* **85,** 724–730. doi:10.23736/S0375-9393.18.12620-4 (2019).

52. Maheshwari, K. *et al.* Performance of the Hypotension Prediction Index with non-invasive arterial pressure waveforms in non-cardiac surgical patients. *Journal of Clinical Monitoring and Computing* **35,** 71–78. doi:10.1007/s10877-020-00463-5 (Jan. 2020).

53. Bhattacharya, S. *et al. A novel classification method for predicting acute hypotensive episodes in critical care* in *ACM BCB 2014 - 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (2014), 43–52. doi:10.1145/2649387.2649400.

54. Ghosh, S. *et al.* Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure. *IEEE Journal of Biomedical and Health Informatics* **20,** 1416–1426. doi:10.1109/JBHI.2015.2453478 (2016).

55. Jiang, D. *et al.* Probability distribution pattern analysis and its application in the Acute Hypotensive Episodes prediction. *Measurement* **104,** 180–191. doi:10.1016/j.measurement.2017.03.030 (2017).

56. Jiang, D. *et al.* A hybrid intelligent model for acute hypotensive episode prediction with large-scale data. *Information Sciences* **546,** 787–802. doi:10.1016/j.ins.2020.08.033 (Feb. 2021).

57. Lee, J. & Mark, R. G. An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. eng. *BioMedical Engineering Online* **9,** 62. doi:10.1186/1475-925X-9-62 (Oct. 2010).

58. Moghadam, M. C. *et al.* Predicting hypotension in the ICU using noninvasive physiological signals. eng. *Computers in Biology and Medicine* **129,** 104120. doi:10.1016/j.compbiomed.2020.104120 (Feb. 2021).

59. Rocha, T. *et al. Wavelet based time series forecast with application to acute hypotensive episodes prediction* in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10* (2010), 2403–2406. doi:10.1109/IEMBS.2010.5626115.

60. Sun, Y. *et al.* Predicting Future Occurrence of Acute Hypotensive Episodes Using Noninvasive and Invasive Features. eng. *Military medicine* **186,** 445–451. doi:10.1093/milmed/usaa418 (Jan. 2021).

61. Solomon, S. C. *et al.* Forecasting a crisis: Machine-learning models predict occurrence of intraoperative bradycardia associated with hypotension. *Anesthesia and Analgesia* **130,** 1201–1210. doi:10.1213/ANE.0000000000004636 (2020).

62. Marsland, S. *Machine learning: An algorithmic perspective* 1–452. doi:10.1201/b17476 (CRC Press, Boca Raton, FL : 2014).

Example of FSW data labelling

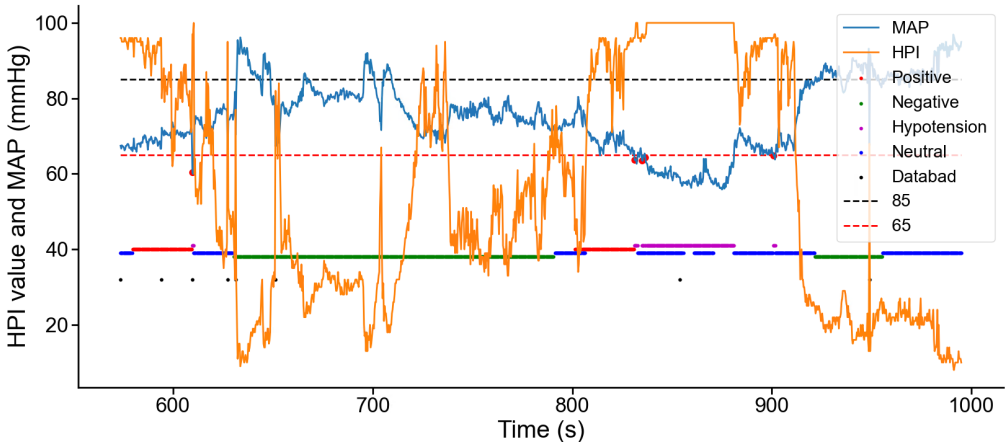


Figure K.1: Forward sliding window: labelling example. Red dots on the MAP curve indicate an onset of hypotension. Colours around the y value of 40 indicate hypotension and labelling of prediction. Purple: Hypotension. Red: predictions on hypotension, i.e. positive points. Green: predictions on non-hypotension, i.e. negative points. Blue: Neutral points, either a leading neutral buffer or a washout period. Black dots indicate data subject to artefacts.

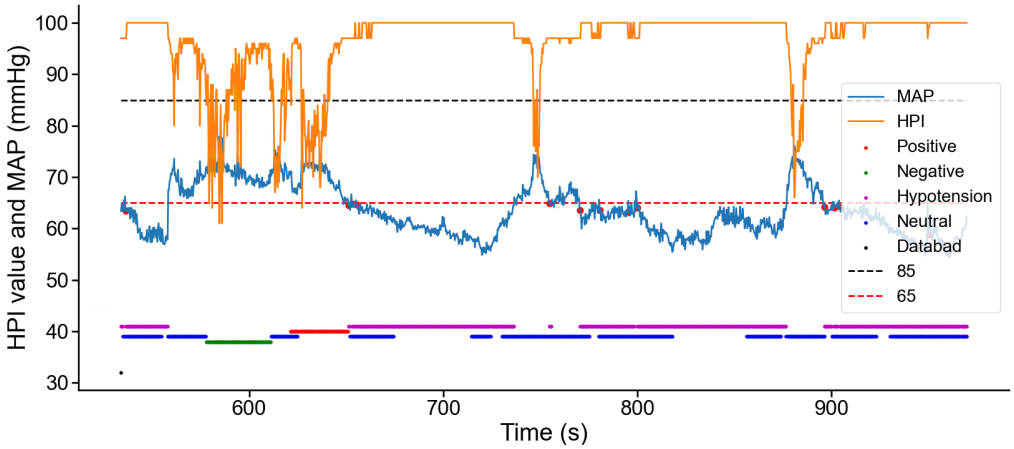


Figure K.2: Forward sliding window: washout example. Consecutive hypotensive episodes are not labelled individually, as the onset of the next hypotensive event is located in the previous washout period. Red dots on the MAP curve indicate an onset of hypotension. Colours around the y value of 40 indicate hypotension and labelling of prediction. Purple: Hypotension. Red: predictions on hypotension, i.e. positive points. Green: predictions on non-hypotension, i.e. negative points. Blue: Neutral points, either a leading neutral buffer or a washout period. Black dots indicate data subject to artefacts.