Predicting patient status dependent on their treatment using a clustering model with SAX

Nina Schoeber

October 14, 2019

Supervisors
prof. dr. A.P.J.M. Siebes
Utrecht University

drs. E. Koomen
University Medical Center Utrecht

dr. T.H. Kappen
University Medical Center Utrecht

ICA-6187285

Contents

List of Figures

Abstract

The goal of this research is to try to predict the condition of a patient in the future, given the current condition and conditional on the treatment using data of the University Medical Center Utrecht. There are two separate use cases: the Pediatric Intensive Care Unit (PICU) and the operating room (OR). In both cases, the haemodynamic parameters are predicted. The prediction is aided by lab measurements and patient information and made conditional on the intervention. The intervention in the PICU dataset consists of inotropes and in the OR dataset it is a combination of inotropes and anesthetics. A model is developed that uses K-Means combined with Symbolic Aggregation ApproXimation (SAX) to cluster the patient windows and uses these clusters and the interventions to build a probability matrix. This probability matrix can be used to predict new cases. The model performs significantly better than a model predicting no change. The model performs equally well as a clustering method using only K-Means, but is better able to consistently cluster the patient status into meaningful categories. The influence of the interventions cannot be isolated as they are too highly correlated with the patient status.

1 Introduction

In an intensive care or operating room, a patient is monitored intensively. There are often a lot of different measurements being taken continuously. Next to the patient's current status, the change in status can provide valuable information about the severity of the status. In order to help the doctors to interpret the measurements and their progression quickly and correctly, research is being done into summarizing or predicting the status of the patient. A prediction of the patient status over time can aid in a quicker assessment of the severity of the situation and thereby help place the focus on the problematic cases and parameters.

The patient status is heavily dependent on the interventions made by the doctors. For example, a blood pressure that is consistently getting lower can suddenly go up because of a change in medication. A prediction model that does not take the intervention into account might learn this pattern and suggest a spike in the blood pressure after any long decrease. This can be avoided by making the dependence on the intervention explicit. It is therefore useful to include the interventions in the prediction model.

The goal of this research is to try to predict the condition of a patient in the future, given the current condition and conditional on the treatment. In a clinical situation, the prediction has to be motivated and interpretable, to be reliable and trustworthy. The second aim is to evaluate the influence of the medical interventions on the prediction.

This leads to two research questions:

1. Can we predict the change in status of the haemodynamic parameters of a patient given the current (and potentially historical) status and the medication they receive?

2. Can we isolate the role of the medication/intervention in this prediction?

The first question is the main goal of this research. If an accurate and transparent prediction model can be established, this can then be used to answer the second question.

1.1 Use cases

Two different clinical environments at the University Medical Center Utrecht are used for this research. While closely related in terms of the variables used, they differ largely on resolution and patient population. This allows for a close comparison of any prediction model in two different situations.

1.1.1 Pediatric Intensive Care: Heart failure

The first case is the Pediatric Intensive Care Unit (PICU). Children in perioperative care for cardiac surgery can have a long stay in the PICU with many interventions. Besides surgery, these children receive different types of inotropes to keep their haemodynamic parameters in a safe range. Therefore, the aim in this use case is to predict the (continuous) haemodynamic parameters of these children. The main parameters of interest in this research are therefore the (arterial) blood pressure, heart frequency and the oxygen saturation of the blood. More information about these parameters follows in Chapter 4. Predicting to what extent the parameters improve given specific inotropes can provide an indication of the severity of the patient's status. In this use case, a low blood pressure, a high heart frequency and a low oxygen saturation are considered problematic.

For the purpose of this research, only the data from children younger than one year old is used. The normal range of the haemodynamical parameters changes significantly with age and large age differences might therefore make a reliable prediction problematic. For example, a large difference in age might cause a clustering algorithm to implicitly group by age instead of by severity. Next to these parameters, discontinuous lab results are available to aid in the prediction.

The available interventions are inotropes. These medications influence the haemodynamic parameters (e.g. increasing the contractility of the heart muscles) and are given through continuous intravenous infusions (IV). The settings of these IV's are available in the dataset for each minute.

In this use case the chosen resolution is hours. This means that given a patient's condition in the current time window of one hour (and potentially historical information), the next hour is predicted. It takes time for the inotropes to take effect and the length of the PICU stay is generally in the order of days or weeks.

1.1.2 Anesthesiology: Haemodynamic fluctuations

The second use case is the operating room. The aim in this use case is to predict the change in haemodynamic parameters during the maintenance phase of the anesthesia in surgery on adult patients. The same haemodynamic parameters were used as in the PICU use case and the same states of these parameters are considered problematic: a low blood pressure, a high heart frequency and a low oxygen saturation.

After the patient has been sedated, the surgery starts and the anesthesiologist maintains the anesthesia by monitoring the haemodynamic parameters and administering sedatives and inotropes.

The maintenance phase was chosen because the induction phase can have many different events and include a large number of single dose injections. In the maintenance phase, pumps are used to continuously administer the medication. The types of pumps and their settings can be varied. All changes in the types of pumps and their speed are recorded in the data. This information can be used to extract the settings of all pumps used for each minute. The data of the maintenance phase of the anesthesia is in form therefore similar to the PICU dataset, as both include continuous medication and the same haemodynamic parameters.

While the end of the anesthesia is not part of the maintenance phase, it is not filtered out. The anesthesia is ended by discontinuing the sedatives. It is in form therefore similar to the maintenance phase.

In this use case the same haemodynamic parameters are to be predicted as in the PICU case but at a much higher resolution. The duration of the surgery is generally in the order of minutes to at most hours and the doses of medication are significantly higher, because they have to take effect quickly in adult patients. Therefore, the chosen resolution is blocks of 6 minutes. The time for medication to take effect is smaller at the higher doses and a divisible time window was needed.

1.2 Problem statement

The goal of this research is to try to predict the status of a patient in the future, given the current status and conditional on the treatment. Given the category in which the patient falls in a time window w_i and the interventions related to this window, we want to predict the category the patient will fall in in w_{i+1}.

Predicting this for each type of treatment available in the dataset shows the influence of the intervention. As the medication is continuously administered, the setting at last minute of the window is defined as the intervention related to that window. Therefore, an intervention I at a given window w_i is defined as the setting of all IV pumps at time $w_i[-1]$, the last minute of w_i. The setting of a single pump iv at time t is defined as iv_t.

$$I_{w_i} = \{\forall_{iv \in IV}\ iv_{w_i[-1]}\}$$

The question then becomes: given the measurement window w_i^p for patient p at time i, can we calculate the probability $c_{w_{i+1}^p}$ for p being in category c in time window $i+1$ for each class $c \in C$, conditional on treatment t_{w_i}?

The probability to calculate is therefore:

$$P(C_{w_i^p} | I_{w_{i-1}^p}, w_{i-1}^p) \tag{1.1}$$

- $p \in P$ patients
- $c \in C$ categories
- $w_i^p \in W$ time windows for patient p at time i
- $I_{w_i^p}$ interventions associated with time window w_i^p

1.3 Background and related work

1.3.1 Time series

The data used for this research consists of time series. Time series data differs from data used in classical machine learning approaches both in its size and nature [15]. Next to a high dimensionality, each time series generally has a high feature correlation and a large amount of noise [3]. This can be seen in the available data as even with the selection of only a few (haemodynamic) parameters, a large number of values are available for each patient and the variables are by definition highly correlated.

This causes a problem when trying to define the similarity of two time series. The phenomenon introduced by [1] as the "curse of dimensionality" states that in high dimensional space, the similarity or distance measure loses its meaning. This occurs because the increase in dimensionality reduces the variation in the distances. The time series become essentially equidistant and therefore similarity or clustering becomes meaningless.

In order to overcome this problem, most data mining methods use some from of dimensionality reduction.

1.3.2 Dimensionality reduction

The goal of dimensionality reduction is to create an approximation of the raw time series that reduces the dimensionality as much as possible without losing the important information. It is not the individual time points that are of interest, but the patterns and trends in the time series. Therefore, the aim is to represent these patterns and trends in a more compact way.

Symbolic Aggregate ApproXimation (SAX) is a discretisation technique proposed by Lin et al. in 2007 as a method to represent time series [10]. It provides dimensionality reduction while still allowing an (lower bounding) approximation of the distance between the original time series. It is essentially an extension to Piecewise Aggregate Approximation (PAA). PAA reduces a time series of n data points to w data points by dividing the time series into w equal sized frames. The vector of the w mean values for each frame is given as the data-reduced representation.

SAX has two extensions to this method. First, the data is normalized before applying PAA. The normalization step converts the data to zero mean and unit variance, by subtracting the mean from the value and dividing the resulting differences by the standard deviation. Next, the vector is discretized by converting the w mean values to a symbol. In order to do this, a breakpoints are calculated such that each symbol is produced with an equal probability. Each of the w values is then checked against the breakpoints and assigned to the corresponding symbol. This yields an ordinal representation which can then be used in place of the original time series.

SAX requires two parameters: the alphabet size a and the number of segments w.

As the symbols can represented by integers representing the order, this representation still allows the use of the Euclidian distance measure. For two equal-sized SAX strings of n symbols (represented by integers), the Euclidian distance is the symmetric straight-line distance through n-dimensional space:

$$d(sax^{(1)}, sax^{(2)}) = d(sax^{(2)}, sax^{(1)}) = \sqrt{\Sigma_{i=1}^{n}(sax_i^{(2)} - sax_i^{(1)})} \tag{1.2}$$

1.3.3 Prediction and clustering

Next to the data preparation, there are many different data mining methods for time series. The most common data mining tasks on time series are clustering and prediction [15].

Clustering divides the data into groups. These groups are defined by the data itself. The clusters are formed by minimizing the distance (or maximizing the similarity) based on some distance measure. There are two main types of clustering: partitional clustering and hierarchical clustering. The most common algorithm used for clustering is K-Means [6]. K-Means is a partitional clustering algorithm that requires the number of clusters K to be given. This algorithm efficiently divides the data into K clusters, finding a (local) minimum from each point in the cluster to the mean. As the algorithm depends on the randomly chosen or preset initial clusters, it can get stuck in a local optimum.

Clustering requires a clear distance measure to minimize. Given any distance measure and a suitable (dimension reduced) representation of the time series, it can find the (locally) optimal set of clusters. This means that it can easily work with representations such as SAX combined with the Euclidian distance.

Clustering is an unsupervised learning method. It does not require any information about any desired labeling. It allows for easy interpretation as labeling or prediction is based on cluster membership, which can be easily explained in terms of distances.

Clustering is not only used as a task in itself, but also as a preprocessing step to pattern discovery algorithms such as time series rule discovery.

Prediction or forecasting attempts to predict a future state. This can be done with many different types of algorithms. A common method is the use of Artificial Neural Networks (ANNs, [19]), but prediction can also be seen as a type of clustering. Instead of generalizing a current state, it uses the information about the cluster members to predict the next state. This can be done by clustering the members based on their current state, and obtaining a generalized future state from the cluster.

2 Motivation

From the related work and the use cases it follows that the model has to be transparent and that some dimension reduction method has to be applied as the data is very high-dimensional.

2.1 Transparency

Methods that work well for time series forecasting are often black-box. For example, ANNs are often used for time series, but these models do not allow for easy rule extraction [19].

In both use cases, it is important that the prediction can be explained by a direct look at the model. Therefore, a prediction through clustering was chosen instead of a neural network approach. In a clustering approach, each prediction can be reduced to the membership of a cluster with the distribution of the future status within that cluster. It therefore allows for a simple rule extraction. These rules can easily be extended with interventions by taking the intersection of cluster membership with the type of intervention. This allows for interpretable (forward) probabilities $P(C_{w_i^p}|I_{w_{i-1}}, w_{i-1}^p)$ as defined in section 1.2.

Another benefit of this choice is that no goal value is required. As this is an unsupervised method, there does not need to be any labeling beforehand. However, it is still possible to label the found clusters later (e.g. by severity) to simplify the interpretation of the results, as the cluster centers or prototypes can be extracted. These prototypes are the mean of all members of the cluster and thereby provide an indication of the general state of the members. This indication can then be used to label the clusters for easier interpretation.

2.2 Dimension reduction and avoiding regression towards the mean

The datasets contain a large number of time windows in which the patient is in a stable state. This is especially the case in the PICU dataset, in which long stable periods can be found. Therefore, a large number of time windows are in a small section of the variable range.

It is therefore expected that a clustering method such as K-Means may focus too much on this center. For example, this could cause time windows outside of this ‘normal’ area to simply be added to the nearest cluster in the center. As the vast majority of the points in the cluster is very close to each other, this might still be a (local) minimum in the sum of squared distances. This would cause the definition of these problematic time windows to change as they are associated with a cluster center in or near the normal range of the variable. On the other hand, it would also change the definition of the center, as the problematic time windows could skew its prototype. For example, when dividing the data into three clusters, the desired division would be low, center and high, with the exact division depending on the variable. With K-Means, the arbitrary splits might cause each of these clusters to contain a part of the center. This changes both the prototypes of the outer categories and the prototype of the center.

However, as we are mainly interested in the time windows that represent a problematic state, these time windows should be identifiable instead of grouped with time windows in a normal range.

Pre-processing the data using SAX not only has the effect of condensing the data, but could also alleviate this problem. As SAX applies symbols on an ordinal scale, this essentially places the values at an equal distance of each other. SAX generates the breakpoints such that each symbol occurs with equal probability. This causes the ‘normal’ area to be broken up into more parts than the problematic values at the edges. This should cause the edge time windows to be grouped together already before the clustering step, which may make it easier for them to become their own cluster instead of being added to the large center cluster.

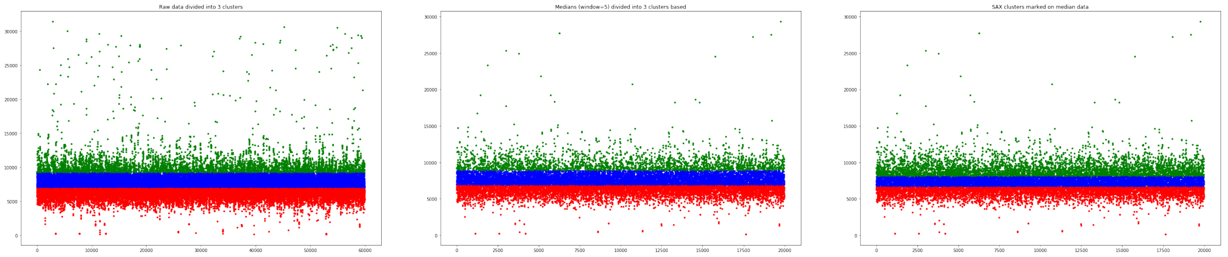
An example of the difference made by including SAX can be found in Figure 2.1. This is an example of the blood pressure, in which we are mainly interested in the lowest category. In Figure 2.1b, the division between the lowest and middle cluster seems arbitrary. The dense region is split into two clusters, while there is no clear change in the data. After SAX is applied with a median filter, it can be seen in Figure 2.1c that this division seems more logical. The dense region is all part of the blue center, while lower red cluster starts where the data becomes less dense.

Figure 2.1c shows that the size of the blue middle cluster is significantly reduced, while still holding the majority of the data. The higher green cluster initially mainly contained outliers, but became an actual category after applying SAX. The percentage of data in each cluster can be found in Table 2.1. While the largest category 1 remains the largest in all three methods, the division becomes more equal.

Figure 2.2 shows the clustering on the reverse transformed SAX levels. This figure shows the equiprobable distribution of the levels.

The alphabet size used for the symbolization should not be too small, as the equidistant breakpoints will then have a negative effect. At a too small alphabet size, the breakpoints only occur in the stable group and the edges are grouped with the closest stable situation, essentially worsening the initial problem.

Depending on the parameters used, the use of SAX may therefore not only reduce the dimensionality of the data, but might also improve the definition of the clusters.



(a) Clustering given by k-means with $k = 3$ on the raw data

(b) Clustering given by k-means with $k = 3$ on the data after applying a median filter

(c) Clustering given by k-means with $k = 3$ on the data after applying SAX

Figure 2.1: The difference made by using SAX on ABP OR data

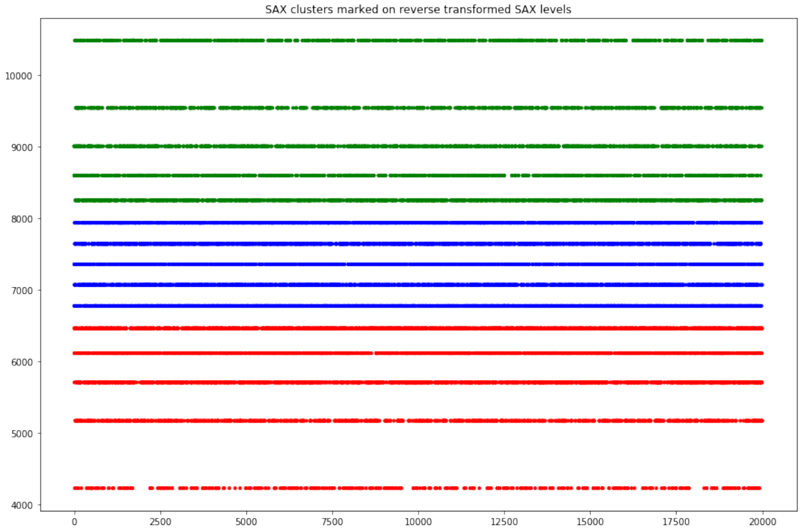


Figure 2.2: The clusters marked on the reverse transformed SAX levels

	0	1	2
Raw data	42.7%	46.7%	10.7%
Medians	39.4%	47.0 %	13.6%
SAX	33.4%	39.5%	27.1%

Table 2.1: Division of data over the three clusters for the methods given in Figure 2.1

2.2.1 Median vs. Mean

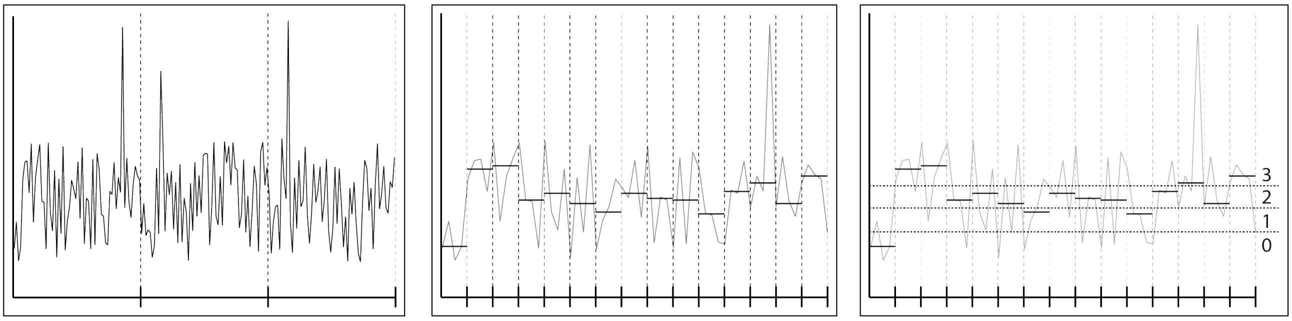
In the original paper by Lin et al. in 2007 in which SAX was introduced, the symbols were assigned using the mean [10]. However, this can be problematic, as the medical data can contain significant outliers. For example, large spikes can occur in the measurements of the arterial blood pressure when blood is drawn using the arterial line (as the line is blocked to the transducer). In order to make the method more robust to outliers, the median is used instead of the mean. The difference in clustering made by applying a median filter can be seen in Figure 2.1. In Figure 2.1a, there are many outliers to be seen. In Figure 2.1b, a median filter with width 5 has been applied and this already reduces the number of outliers that are directly visible. Table 2.1 shows that this reduction of the influence of outliers changes the division of the data over the clusters. The percentage of data in both the highest and lowest category has changed. This shows that the outliers may influence the division of the data. Therefore, the median was chosen for the SAX method to make it more robust.

3

Model description

3.1 Data preparation

The first step of the model is to prepare the data for training. In this preparation step, the raw data is transformed to a list of categorized time windows. Each variable is categorized separately. The final category of each time window is determined as the intersection of the variable categories.



(a) Windowing the raw data

(b) Dividing the window into segments of equal size and taking their median

(c) Applying SAX to the medians

Figure 3.1: The pipeline of the data preparation

3.1.1 Windowing

First, the data is divided into W time windows (Figure 3.1a). The size of each window $w \in W$ and the step between the start times of subsequent windows are given as a parameter for each use case. The step between two subsequent windows w_i and w_{i+1} is chosen smaller than the window size, so that subsequent windows partially overlap. This is done to ensure that the exact timing does not influence the prediction. For example, a window with a large spike at the start has a large distance to a window with a similar spike at the end. However, this does not necessarily mean that the spike has a different meaning, as the location in the window depends on an arbitrary start time. Using a sufficiently large overlap, both windows are included in the dataset.

If a window misses data for any minute, it is discarded.

3.1.2 Symbolic Aggregate Approximation

The next step is to transform each window into a list of symbols using SAX as described in Section 1.3.2.

Next to the data, the SAX method requires two parameters: n and a. The parameter n determines the length of the dimension-reduced representation that is produced by SAX. The parameter a represents the alphabet size, which determines the number of symbols that are used in the discretized representation of the dataset.

First, each window w_i is subdivided into n equal sized subsegments. The median of each segment $s_j \in w_i$ is calculated and a vector of medians $M = m_1, ..., m_n$ is produced, as can be seen in Figure 3.1b.

Next, breakpoints $B = \beta_1, ...\beta_{a-1}$ are determined such that it assigns a symbols with equal probability. Each value m_i in the vector of medians is then checked against the breakpoints and assigned to the corresponding symbol, as can be seen in Figure 3.1c. The symbols were chosen as integers (in ascending order) for easier distance calculation in the clustering step. The assignment is done as follows:

$$\hat{m}_i = j, \qquad \text{iff} \quad \beta_{j-1} \leq m_i < \beta_j$$

This results in a list of n symbols for each window.

3.1.3 Clustering

Next, these vectors of SAX symbols are fed to a K-Means clustering algorithm. The initial cluster centers were randomly selected by the algorithm. In this algorithm, the Euclidian distance is used to divide the set of SAX vectors into $k = 3$ clusters. Each cluster receives a label $C = 0, .., k-1$, which is then defined as the category for all of its members for that variable. The labels are applied such that the order of the cluster labels corresponds with the ascending order of the cluster centers.

It would also have been possible to cluster the data after combining the variables. However, combining multiple SAX strings would significantly increase the dimensionality of the data that is to be clustered. By clustering each variable separately, this dimensionality is limited and it is therefore easier to get meaningful, interpretable clusters.

3.1.4 Combining the variables

Performing the aforementioned steps for each variable results in a dataframe with a row for each time window containing a category for each variable. The intersection of these variable categories is then used as the overall category for that time window. If any variable has no category for a certain window, that window is discarded. An example of variable categories and their resulting overall category can be seen in Table 3.1. A graphical representation of the category division can be found in Figure 3.2.

	Start time	ABP	HF	SAT	Category
0	0	0	2	0	$ABP0 \cap HF2 \cap SAT0$
1	60	1	2	0	$ABP1 \cap HF2 \cap SAT0$
2	120	1	2	0	$ABP1 \cap HF2 \cap SAT0$
3	180	0	2	1	$ABP0 \cap HF2 \cap SAT1$
4	240	0	2	1	$ABP0 \cap HF2 \cap SAT1$

Table 3.1: Example of category combination

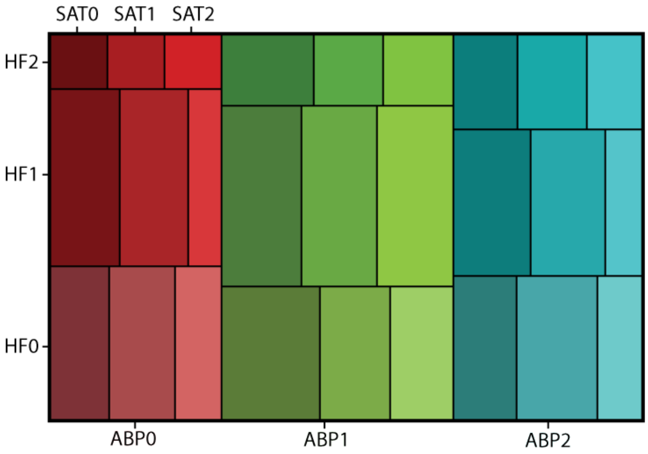


Figure 3.2: The resulting category division

3.1.5 Next and previous categories

Next, the window to predict and its category need to be found. For a current window with start time $st_{current}$ and Δt being the step size between two consecutive windows, the window to predict can be any window such that its start time $st_{predict} = st_{current} + k * \Delta t$ for any positive integer k. For the purpose of this research, the next window is chosen as the time window that starts the minute after the current window ends (i.e. $st_{predict} = st_{current} + windowsize$).

It can also be useful to add history to the model. Therefore, the previous categories need to be found as well. The start time for each (positive integer) level of history h is defined as follows: $st_{history=h} = st_{current} - h * windowsize$.

3.1.6 Covariables

Covariables were used to aid the prediction. In the current and previous categories, these variables are added to the intersection. This means that adding covariables further divides each area in Figure 3.2. In the category to predict, covariables are not included and that category division therefore remains unchanged. This ensures that they add information without increasing the number of prediction choices. The variables are discretized using the same method and parameters as the prediction variables, with the exception of some pre-categorized information that is added in the OR dataset.

As the covariables in the PICU dataset consist of discontinuous measures such as lab values, these measurements are not present in every window. Only taking windows with values for all variables would therefore cause a significant loss of data points and introduce bias. This is handled in two different ways. First, all lab values in the PICU dataset are linearly interpolated with a maximum distance of one day. If there is more than one day between two measurements, there are no values between them. This absence might also add useful information as the choice of whether or not a blood test is performed can provide information about the patient status. Therefore, the absence of a measurement is also added as a category. For example, the covariable blood lactate (LAC) can have the following values (for $k = 3$): $[LAC0, LAC1, LAC2, LAC-]$, with $LAC-$ representing the absence of a measurement.

3.1.7 Merging

The interest is mainly in the problematic 'side' categories (low ABP, high HF, low SAT). If a patient's status is currently within a normal range for all variables, the exact status within this range is not important. Likewise, movements within this normal range are not of interest. Therefore, a possible simplification step to the model is to merge the center categories. This is done by taking all categories that do not have at least one category from $[ABP0, HF2, SAT0]$ (the side categories for $k = 3$) and combining them into a large category $Center$. Combining ABP, HF and SAT with three clusters each, causes the observations in the categories marked in Figure 3.3 to be combined in the $Center$ category. Covariables do not change which categories are merged and retain their original value in the current and previous windows (e.g. $ABP1 \cap HF1 \cap SAT1 \cap LAC0$ becomes $Center \cap LAC0$). Therefore, there can still be variation in the categories in the center if their covariables vary. As mentioned in section 3.1.6, covariables are not included in the prediction category. Therefore, a prediction category with no variable in $[ABP0, HF2, SAT0]$ simply gets the category $Center$.

This merging removes some details and can therefore not be applied when the exact value is to be predicted. However, when focus is on the change in categories, this does not remove any important information. In order to accommodate this difference, the model can be used both merged and unmerged.

3.1.8 Intervention

The last step of the data preparation is to add the interventions to the dataset. The intervention doses are manually categorized into levels in order to allow for easier clinical interpretation. As continuous medication is used, the intervention is taken as the value at the last minute of the time window. The intervention supports the move from one category to another and may be influenced by the status in the current window. For example, a medication dose can be increased because of a decline in patient status. The interaction between the medication and the status should then in turn influence the status in the next time window.

3.1.9 Result

The result of the data preparation is a data frame with for each row the current (and potentially previous) category, the prediction category and the intervention as values to be used for prediction. Beside these prediction columns, the data frame also has the patient (pseudo) ID, start time and a reference to the next category. These are not used for the prediction, but are present for the purpose of verification and retrieval of the original window for the calculation of performance measures.

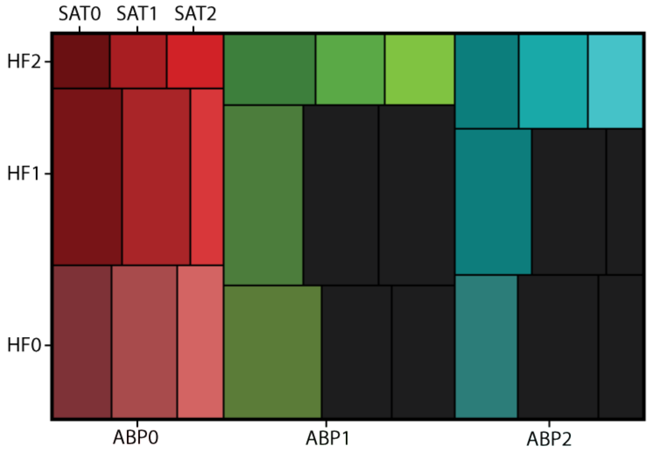


Figure 3.3: The category division with the merged categories marked

The data frame for the example patient in Table 3.1 with a history of $h = 1$, covariable LAC and added interventions can be found in Table 3.2. As this example only has information for five consecutive hours, only three rows have information for both the previous and next hour. The last step of the data preparation is therefore to remove all rows with missing data in one of the category columns. In this example, the greyed out first and fifth row are therefore removed. In the actual datasets, there are many more consecutive windows. Therefore, this last filter only removes a small number of windows near the start, end and gaps in the data for each patient.

The final data frame can then be used for training and testing. The dataset is randomly split into a training and testing set (as further discussed in Chapter 5). The training part uses all categories with the interventions to build a probability matrix. The test part then uses the current (and potential previous) category and the intervention combined with the probability matrix to predict the next category. This predicted category can then be compared to the actual next category from the dataframe.

ID	st	Interv.	Category	Next category	Category History h_1
1	0	Adr_{Low}	$ABP0 \cap HF2 \cap SAT0 \cap LAC0$	$ABP1 \cap HF2 \cap SAT0$	-
1	60	Adr_{Low}	$ABP1 \cap HF2 \cap SAT0 \cap LAC1$	$ABP0 \cap HF2 \cap SAT1$	$ABP0 \cap HF2 \cap SAT0 \cap LAC0$
1	120	None	$ABP1 \cap HF2 \cap SAT0 \cap LAC1$	$ABP1 \cap HF2 \cap SAT1$	$ABP1 \cap HF2 \cap SAT0 \cap LAC1$
1	180	None	$ABP0 \cap HF2 \cap SAT0 \cap LAC1$	$ABP1 \cap HF2 \cap SAT1$	$ABP1 \cap HF2 \cap SAT0 \cap LAC1$
1	240	None	$ABP0 \cap HF2 \cap SAT1 \cap LAC1$	-	$ABP0 \cap HF2 \cap SAT0 \cap LAC1$

Table 3.2: An example of a data frame with no merged categories and parameters $h = 1$, $k = 3$, $windowsize = 60$

3.2 Training

In the training step of the model, the probability matrix is built. This matrix contains a cell for each combination of current hour $w_i \in W$ and intervention $t \in T$. If previous time windows (history) are included in the prediction, they are represented by extra dimensions of the matrix. The cells are then split based on the category in those previous windows.

Each cell of the probability matrix contains a dictionary, which is a set of key-value pairs. Initially, this dictionary is empty. The training algorithm processes each row in the training dataset by finding the correct cell in the probability matrix based on the category, intervention type and history. The given next category is then added as a key to the dictionary, with a value of 1. If it is already present in the dictionary, the value is incremented. This results in a count of how often each possible next category occurs for each specific start situation.

After all rows in the training dataset have been processed, each dictionary is divided by its sum. This means that all counts are replaced by the proportion. When the training and test dataset have the same distribution, this proportion can be seen as the probability. In the testing phase of this research, the training and test sets are randomly sampled from the dataset and therefore can be assumed to have the same distribution.

This yields a probability matrix with for each combination of previous windows the probability $P(C_{w_i^p} | I_{w_{i-1}^p}, w_{i-1}^p)$ as defined in section 1.2.

Using the example categories from Table 3.2 and adding some extra data, an example probability matrix can be found in Table 3.3. The category with the highest probability is marked as bold. This example does not include history. Each level of history is added as an extra dimension.

Current / Intervention	$ABP0 \cap HF2 \cap SAT0 \cap LAC0$	$ABP0 \cap HF2 \cap SAT0 \cap LAC1$	$ABP1 \cap HF2 \cap SAT0 \cap LAC1$
None	$\{\boldsymbol{ABP0 \cap HF2 \cap SAT0}$:0.8, $ABP1 \cap HF2 \cap SAT0$:0.2$\}$	$\{\boldsymbol{ABP0 \cap HF2 \cap SAT0}$:0.9, $ABP1 \cap HF2 \cap SAT0$:0.1$\}$	$\{\boldsymbol{ABP0 \cap HF2 \cap SAT0}$:0.7, $ABP1 \cap HF2 \cap SAT0$:0.3$\}$
Adr_{Low}	$\{ABP0 \cap HF2 \cap SAT0$:0.25, $\boldsymbol{ABP1 \cap HF2 \cap SAT0}$:0.75$\}$	$\{\boldsymbol{ABP0 \cap HF2 \cap SAT0}$:0.6, $ABP1 \cap HF2 \cap SAT0$:0.4$\}$	$\{ABP0 \cap HF2 \cap SAT0$:0.1, $\boldsymbol{ABP1 \cap HF2 \cap SAT0}$:0.9$\}$

Table 3.3: An example of the probability matrix without history

3.3 Prediction

The resulting probability matrix can be used to predict new cases. The prediction function finds the correct cell in the probability matrix based on the category, intervention type and potential history and retrieves the dictionary. It then selects the key with the highest value. This is the category with the highest probability $P(C_{w_i^p}|I_{w_{i-1}^p}, w_{i-1}^p)$ of being the next category associated with this time window. This key is returned as the prediction for the next category of the new time window.

4

Data description

4.1 Datasets

The two use cases each have a separate dataset. While the haemodynamic parameters used in the datasets are the same, their distribution differs greatly. There is also a different set of covariables for both datasets.

4.1.1 Pediatric Intensive Care

The PICU dataset consists of 2250 patients from 2016 to 2018. Only the data of patients younger than one year old was used. Of these patients, 88 have at some point received inotropes and had an arterial line. The final dataset consists of the data for these 88 patients over a period of one month.

This dataset is then divided into time windows by using a rolling window with a width of 60 minutes and a 10 minute step. This means that each pair of consecutive windows has a 50 minute overlap. This leads to a total of 97102 one-hour windows with data for each minute for all three variables. Of these windows, 86046 have a next hour to predict.

Finally, there are 75567 time windows that have a next hour and two hours of history.

4.1.2 Anesthesiology

The OR dataset consists of 9766 adult patients from July 2014 to July 2019. Of these 9766 patients, 6798 had an arterial line. The data was filtered such that only the maintenance (and end) phase of the anesthesia is used. This was done by taking the time given as the end of the induction phase for each patient and only taking the data after this time.

This dataset is then divided into time windows by using a rolling window with a width of 6 minutes and a 3 minute step. This means that each pair of consecutive windows has a 3 minute overlap. This leads to a total of 476610 time windows with data for each minute for all three variables. Of these windows, 456053 have a next window to predict.

Finally, there are 422802 windows that have a next window and two previous six-minute blocks of history.

4.2 Prediction variables

Both datasets contained a large number of variables with different degrees of abundance and importance. Therefore, only the variables that were selected for this research are outlined below. The research focused on three haemodynamic parameters: arterial blood pressure, heart frequency and oxygen saturation.

The data is marked according to the Early Warning System (EWS) as used in the University Medical Center Utrecht, or other reference values from the use cases when available. For the PICU dataset, the Pediatric Early Warning System (PEWS) is used. In PEWS there are different scores for children from 0 to three months old and from three months to a year old. As the majority of the PICU dataset is younger than three months old (as shown in Figure 4.9), the reference values from this age group are used. A higher score corresponds with a higher risk. The data is marked to show the distribution of values with regard to their clinical interpretation.

4.2.1 Arterial blood pressure

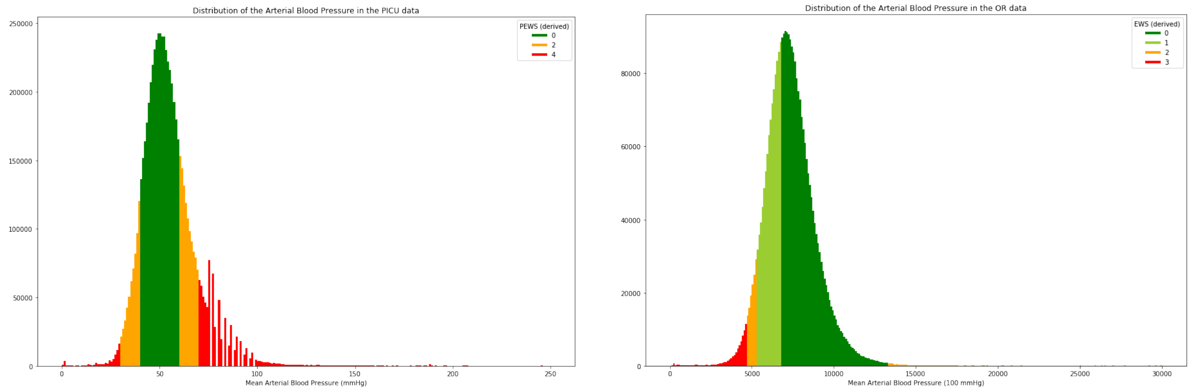
The arterial blood pressure (ABP) is a continuous measurement in both use cases. Whenever a patient has an arterial line, the blood pressure is recorded every minute. The blood pressure is measured through the arterial line and the systolic, diastolic and mean pressure are recorded. The systolic blood pressure ABP_{sys} is the maximum pressure that occurs when the left ventricle contracts. The diastolic blood pressure ABP_{dia} is the minimum pressure between two contractions. The mean blood pressure ABP_{mean} is calculated using the following formula [4]:

$$ABP_{mean} = \frac{ABP_{sys} + (2 * ABP_{dia})}{3}$$

For the purpose of this research, the mean blood pressure is used as prediction variable as it was available in both datasets. The normal range of the blood pressure increases with age.

The blood pressure is recorded in mmHg. In these use cases, the blood pressure is considered problematic when it is too low, as this suggests insufficient blood flow.

As can be seen in Figure 4.1, the blood pressure is relatively normally distributed. In the PICU dataset, the mean ABP_{mean} is 54.3 mmHg and the median ABP_{mean} is 52.0 mmHg. In the OR dataset, the mean ABP_{mean} is 73.5 mmHg and the median ABP_{mean} is 72.0 mmHg.



(a) Distribution of ABP_{mean} in the PICU dataset (b) Distribution of ABP_{mean} in the OR dataset

Figure 4.1: The distribution of the arterial blood pressure in both datasets

4.2.2 Heart frequency

The heart frequency (HF) or pulse of a patient is also available on a minute by minute basis in both datasets. It is measured as the number of contractions of the heart per minute and is therefore recorded in bpm. The heart frequency depends on the patient's age and condition. For example, a young child has a higher heart rate than an older patient. A patient with fever or heart failure generally also has a higher heart rate.

The cardiac output is determined by the volume of blood that is being displaced in one heartbeat (stroke volume) and the heart frequency [8]. A neonate can only vary its heart rate to change its cardiac output. Older children and adults can vary both their heart rate and their stroke volume.

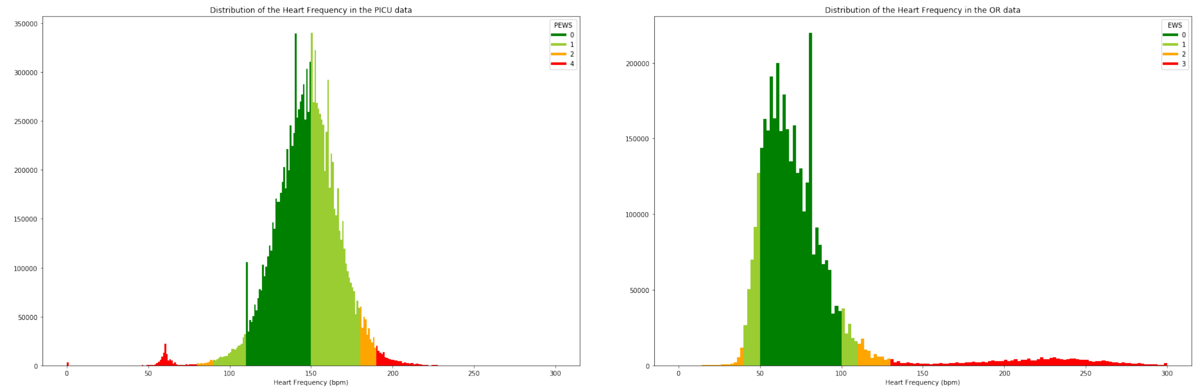
In these use cases, the heart rate is considered problematic when it is too high.

As can be seen in Figure 4.2, the average heart rate is significantly higher in the PICU dataset than in the OR dataset. In the PICU dataset, the mean HF is 146.2 bpm and the median HF is 147 bpm. The heart frequency in the PICU dataset seems to be relatively normally distributed, whereas the OR dataset has a larger percentage of values on the high end of the distribution. In the OR dataset, the mean HF is 77.2 bpm and the median HF is 68 bpm.

In both datasets, the majority of the values lies in the range corresponding to (P)EWS score 0 or 1.

4.2.3 Oxygen saturation

The oxygen saturation (SAT) is the percentage of the hemoglobin in the blood that is oxygen-saturated. This should be 95-100% when the blood leaves the lungs. This is the arterial saturation, or SaO2. The saturation can be approximated using the peripheral saturation SpO2 [7]. This is measured using a pulse oximeter at the



(a) Distribution of HF in the PICU dataset

(b) Distribution of HF in the OR dataset

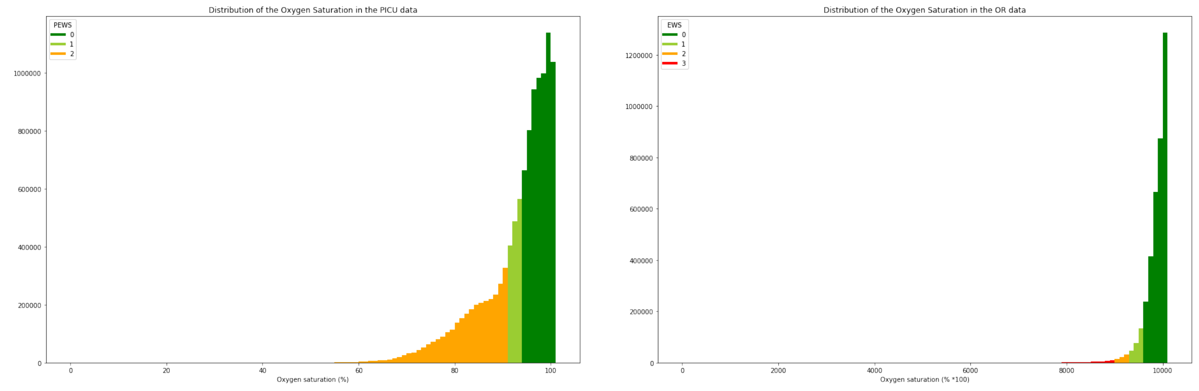
Figure 4.2: The distribution of the heart frequency in both datasets

finger, which measures the saturation based on the difference in absorption of red and infrared light by the oxygenated hemoglobin and deoxygenated hemoglobin. In a healthy adult, this should be 95-100%, with levels below 90% considered hypoxemia, insufficient oxygen in the blood.

In the PICU use case, significantly lower values for oxygen saturation occur, as these children often have some form of heart failure. Within the population of these patients with a heart disorder, there are two cases for the saturation. Patients with two ventricles and no mixing should have normal saturation levels of 95% to 100%. Patients with only one ventricle or two with blood mixing between them, the saturation level should be between 75% and 85%. If such a patient has a saturation of $SpO2 > 85\%$, this can mean that there is too much blood going through the lungs and not enough through the body, which can cause a heart failure situation and anaerobe lactate production.

Figure 4.3 shows that the PICU dataset contains a lot more lower saturations than the OR dataset. In the PICU dataset, the mean SAT is 92.9% and the median SAT is 95.3%. In the OR dataset, the mean SAT is 98.2% and the median SAT is 99%. This difference occurs because of the potential heart disorders mentioned above.

A low saturation is problematic as this can lead to hypoxia, insufficient oxygen in the tissue.



(a) Distribution of SAT in the PICU dataset

(b) Distribution of SAT in the OR dataset

Figure 4.3: The distribution of the oxygen saturation in both datasets

4.3 Covariables

Next to the prediction variables, covariables are added. These variables supply info but are not predicted themselves. There are two types of covariables: dynamic and static. The value for the dynamic variables can differ between the current hour and potentially included history, while the static information is given from a preoperative survey (POS).

Covariables are only used in the category of the current hour and history, not in the category to predict. As the static information does not change between two time windows, this is only added at the current hour.

In the PICU dataset, the covariables lactate, creatinine and ASAT are used. Others were explored, but these proved to be the most informative in initial runs. These are all dynamic, discontinuous lab measurements.

In the OR dataset, the dynamic variable expirated vapor is used and the static information age group, BMI class, ASA score, specialism and hypertension.

4.3.1 Lactate

Lactate is a product of anaerobic metabolism. When there is not enough oxygen available for the cells, they can partially convert to anaerobic metabolism to produce energy. The lactate is measured in the lab and is therefore a discontinuous measure. It is measured more often for at risk patients. Therefore, the frequency at which the lactate level is measured can be an indication of the physician suspecting heart failure. While higher lactate values are an indication of the patient's risk, the lactate clearance may provide more information [9]. Therefore, multiple consecutive measurements of lactate might provide more information than the absolute value.

Normal lactate levels are 0-2 mmol/L. At a higher level, the anaerobic metabolism has started. Lactate levels of 2-5 mmol/L are considered reasonably high, but can be normal after surgery. Lactate levels of above 5 are considered high.

The mean value for lactate in the PICU dataset is 1.58 mmol/L and the median value is 1.4 mmol/L. The distribution of lactate values can be found in Figure 4.4. Most of the data is in the normal range. Only a very small number of lactate values in the dataset are higher than 5 mmol/L.

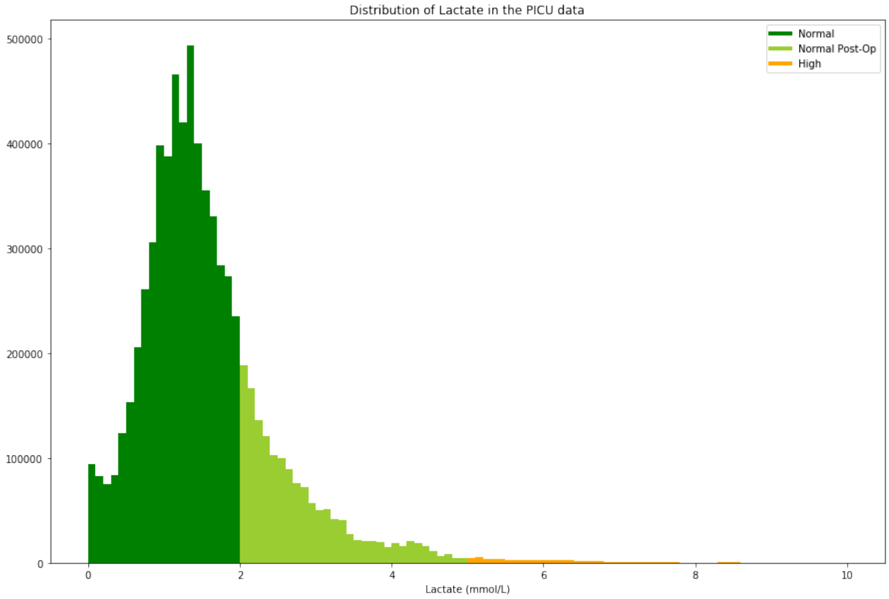


Figure 4.4: Distribution of Lactate in the PICU dataset

4.3.2 Creatinine

Creatinine is a waste product produced by the muscles. This is filtered out of the blood by the kidneys. A high level of creatinine in the blood can therefore indicate renal dysfunction.

In patients with heart failure, renal dysfunction occurs often [11]. The heart failure may be worsened by the renal dysfunction and vice versa. Therefore, the renal function can be an important indicator of the patient's status.

Normally, the creatinine concentration drops in the first year and then increases with age. This is because a neonate's creatinine levels are influenced by the creatinine levels of the mother. The mean value for creatinine in the PICU dataset is 51.6 μmol/L and the median value is 38.6 μmol/L. The distribution of the creatinine levels can be found in Figure 4.5.

4.3.3 ASAT

ASAT (aspartate aminotransferase) is an enzyme mainly found in the liver and muscles, such as the heart muscle [17]. It is closely related to ALAT (alanine aminotransferase), which was also present in the dataset, but which was not used as it had little impact on the prediction when used alongside ASAT.

A high level of ASAT can indicate damage to the liver cells or muscles, for example because of insufficient perfusion. The mean value for ASAT in the PICU dataset is 75.1 U/L and the median value is 40.4 U/L. The distribution of the ASAT values can be found in Figure 4.6.

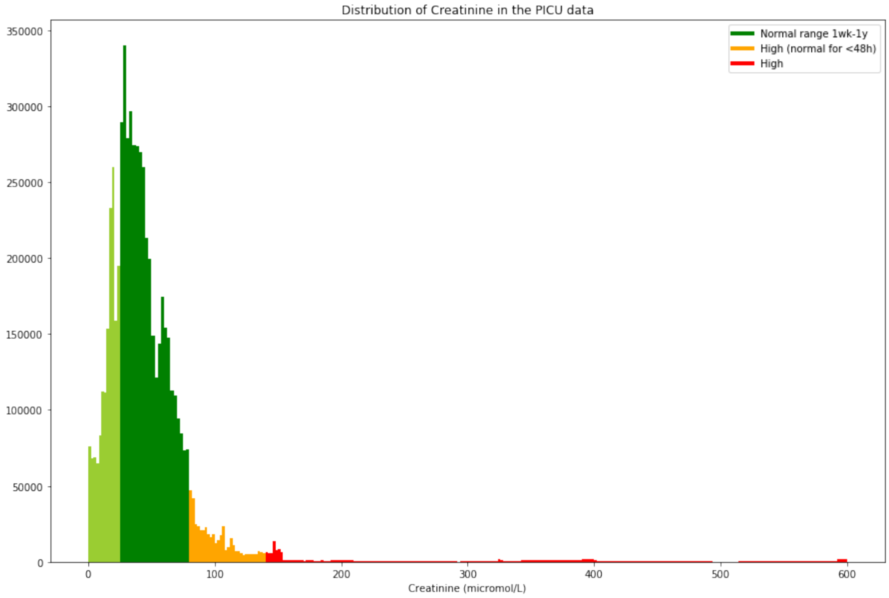


Figure 4.5: Distribution of Creatinine in the PICU dataset

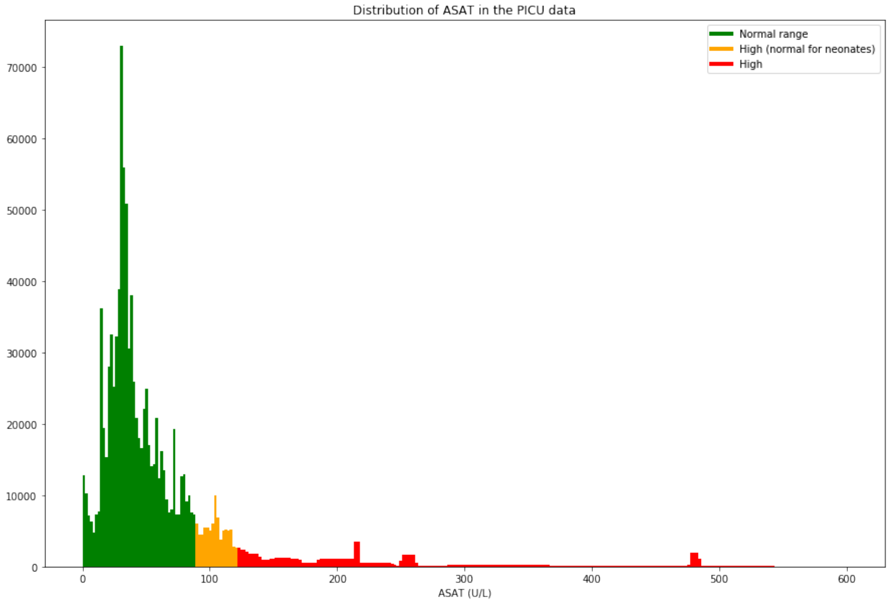


Figure 4.6: Distribution of ASAT in the PICU dataset

4.3.4 Vapor

The concentration of expirated (end-tidal) vapor is an indication of the concentration of inhaled sedatives in the blood. It is the direct result of an intervention, but it is added as a covariable because it also gives an indication about the patient's status.

There are two types of vapor in the dataset: isoflurane and sevoflurane. These have similar effects at different concentrations. In order to use them together, the values need to be normalized. This can be done using the Minimum Alveolar Concentration (MAC) [12]. The MAC is the concentration at which 50% of the patients does not respond to a surgical incision. The ratio between the MAC for isoflurane and sevoflurane is 1.54. By multiplying all concentrations for isoflurane by 1.54, the end-tidal vapor is normalized to sevoflurane.

In the normalized dataset, the mean value for expirated vapor in the OR dataset is 1.21 vol% and the median value is 1.22 vol%.

The distribution of the expirated vapor can be found in Figure 4.7.

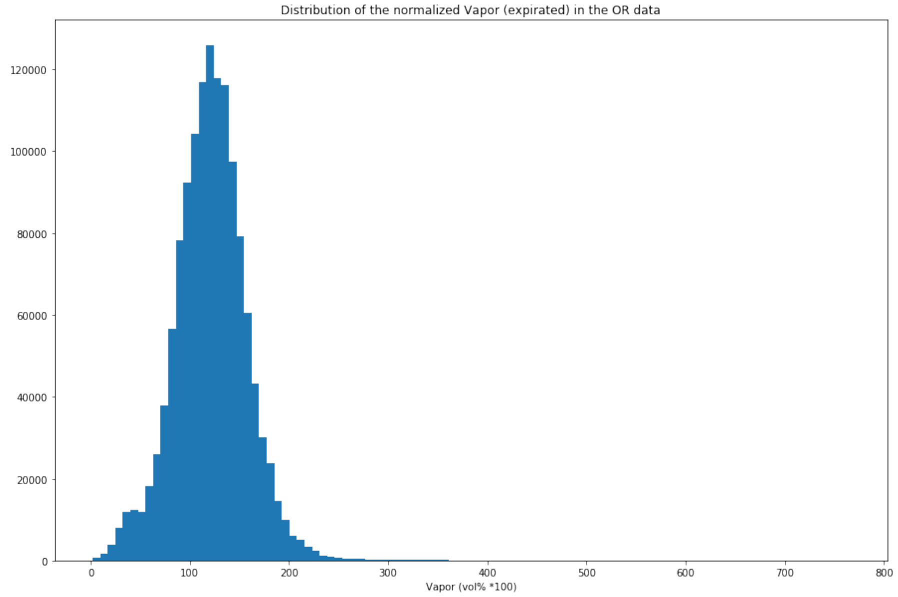


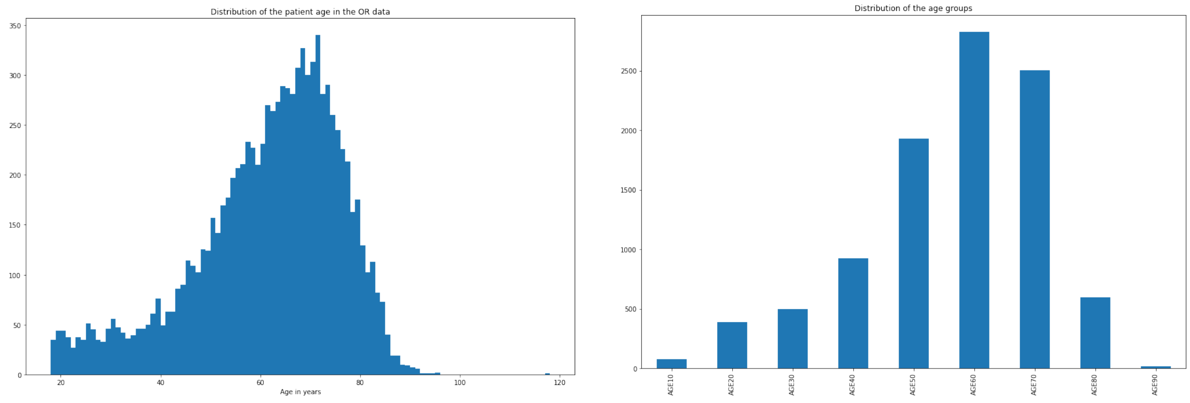
Figure 4.7: Distribution of Vapor in the OR dataset

4.3.5 Age group

The age of a patient can have a large influence on the patient status. For example, the normal range of the haemodynamic parameters differs greatly between the young children and older patients. But also within the adult dataset, the age might give some extra information about the patient's health.

In the PICU dataset, data of children older than one year old have been filtered out. The mean age is then 82.4 days old and the median age is 56 days old. The distribution of ages within this group of children younger than one year old can be found in Figure 4.9.

As the PICU dataset is already filtered by the patient's age, it is only added as a covariable in the OR dataset, where it is included as static information. The mean age in the OR dataset is 60.9 years old. The median age in the OR dataset is 64.0 years old. It was chosen to divide the age into ten year groups, in order to reduce the variability, while keeping most of the information. The distribution of the age (groups) can be found in Figure 4.8.



(a) Age in years

(b) Age groups

Figure 4.8: Distribution of age in the OR dataset

4.3.6 BMI class

A severely under- or overweight patient might respond differently to a medication or status change than a patient at a normal weight. Therefore, the BMI class was added as a covariable to the OR dataset. The BMI is calculated using the formula $\frac{patientweight(kg)}{(patientheight(m))^2}$. The BMI was then categorized into categories underweight, normal weight, overweight and obese [16].

The distribution of the BMI groups in the OR dataset can be found in Figure 4.10.

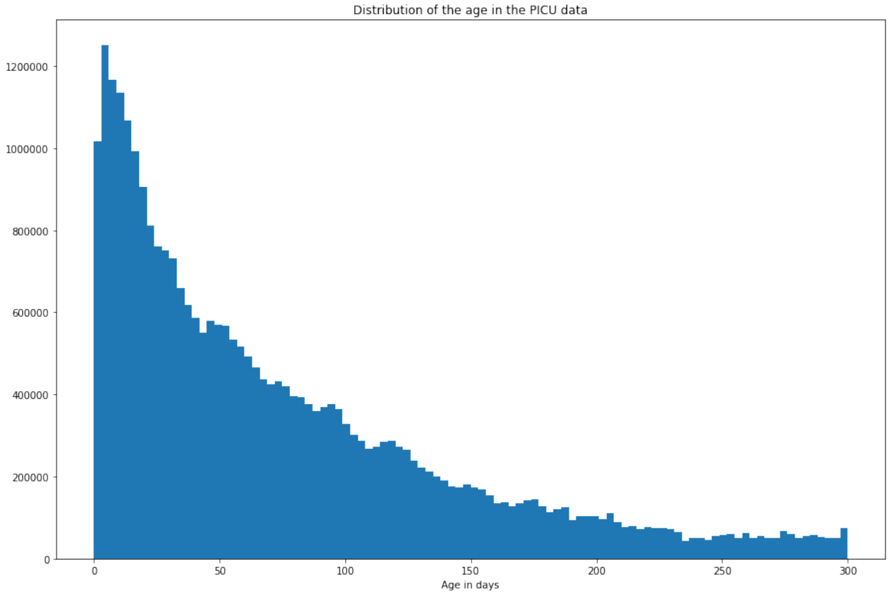


Figure 4.9: Distribution of age in the PICU dataset

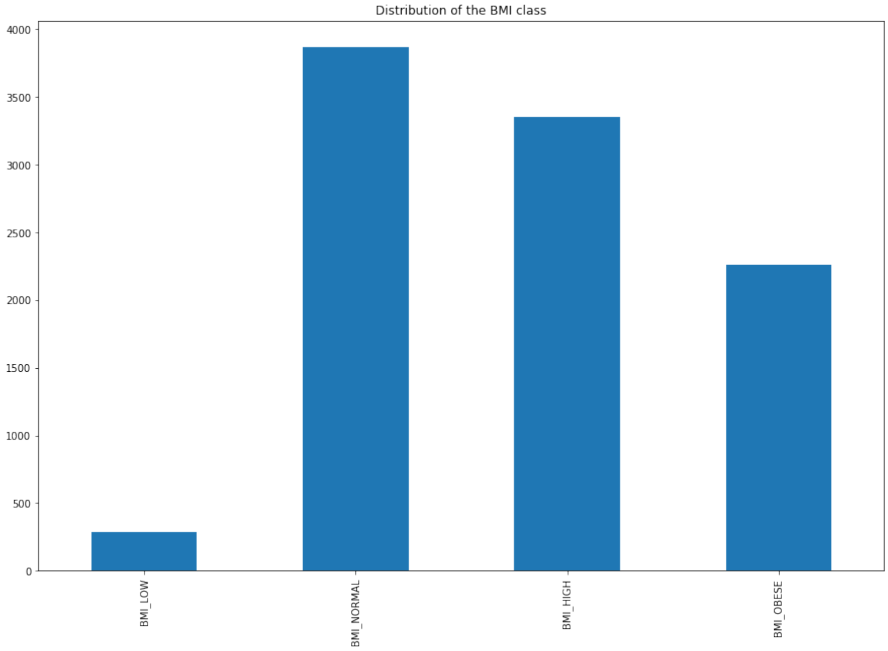


Figure 4.10: Distribution of BMI in the OR dataset

4.3.7 ASA score

The American Society of Anesthesiologists' (ASA) Physical Status Classification is a score used to classify the physical status of a patient before surgery [14]. The classification can be found in table 4.1.

Class	Physical status
1	A normal healthy patient
2	A patient with mild systemic disease
3	A patient with severe systemic disease
4	A patient with severe systemic disease that is a constant threat to life
5	A moribund patient who is not expected to survive without the operation

Table 4.1: ASA classification (from [13])

If the surgery is an emergency, the letter E is added to the class. Class 6 has later been added to classify patients that are declared brain-dead, where the surgery is for donor purposes [13]. For the purpose of this research, data from patients with an E-class or class 6 are not included.

The distribution of ASA scores in the dataset can be found in Figure 4.11. Most patients have ASA score 2 or 3. There is also a large group of patients for which no ASA score is recorded.

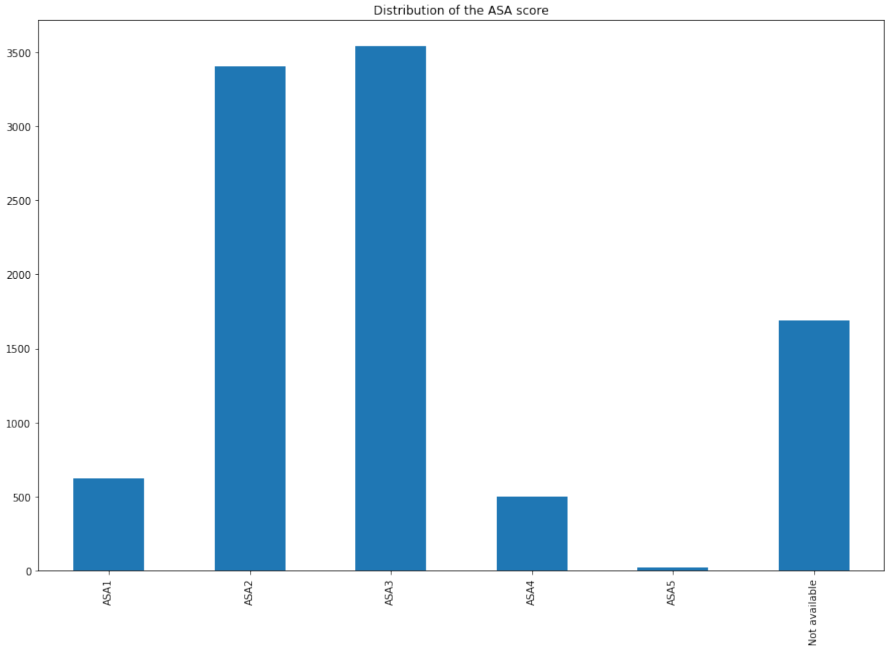


Figure 4.11: Distribution of the ASA score in the OR dataset

4.3.8 Specialism

The medical specialism of the surgery can also provide insight in the status of the patient. The different specialisms are represented by three-letter abbreviations. The most common specialism in the dataset is cardiothoracic surgery (CTC). This specialism includes heart and lung surgery. It is therefore highly related to the haemodynamic parameters.

The distribution for the different medical specialisms can be found in Figure 4.12.

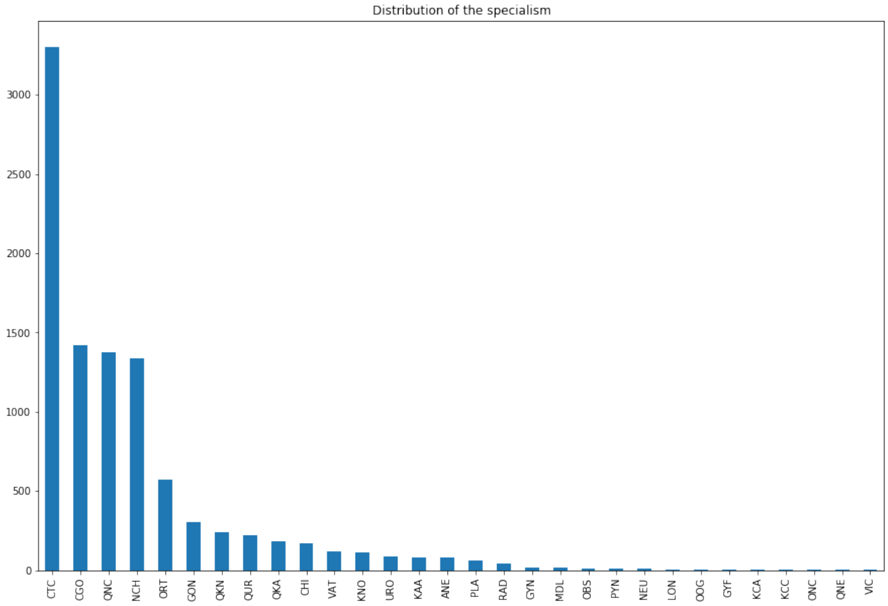


Figure 4.12: Distribution of medical specialisms in the OR dataset

4.3.9 Hypertension

Patients with hypertension have a chronic high blood pressure. Hypertension can indicate a health risk and can skew the normal ranges for the blood pressure.

Hypertension is represented by a yes/no question in the POS. There are also many cases in which the answer is missing. The distribution for hypertension can be found in Figure 4.13.

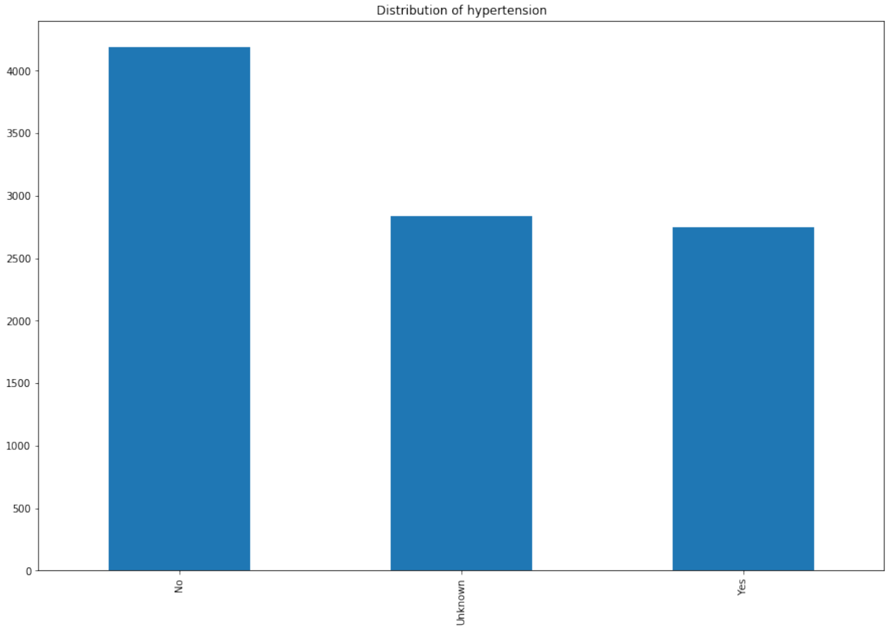


Figure 4.13: Distribution of hypertension in the OR dataset

4.4 Interventions

In the PICU-dataset, the available interventions are six types of inotropes: adrenaline, noradrenaline, dopamine, dobutamine, milrinone and phenylephrine. As milrinone is generally on the same setting and phenylephrine barely occurred in the dataset, these were excluded. In the OR-dataset, the available interventions are a combination of inotropes and anesthetics. As this was a very large set with varying frequencies, this was narrowed down to four types of intervention: phenylephrine, propofol, remifentanil and norepinephrine (noradrenaline).

All of these medications are given through continuous intravenous infusions (IV).

In total, this yields a list of seven intervention types: five types of inotropes and two sedatives/analgesia. In the Table 4.2, the effects of the medications and the dataset in which they are present are outlined briefly. The inotropes have been manually discretized into levels based on the levels used in practice in the different use cases.

	PICU	**OR**	**Effect**
Inotropes			
Adrenaline	X		HF↑ ABP↑
Noradrenaline	X	X	ABP↑
Dobutamine	X		HF↑ ABP−|↓
Dopamine	X		HF↑ ABP↑
Phenylephrine		X	ABP↑
Anesthetics			
Propofol		X	Sedative ABP↓
Remifentanil		X	Analgesic ABP↓ HF↓

Table 4.2: The different types of interventions

4.4.1 Interventions PICU

The distribution of the combinations of inotropes in the PICU dataset can be found in Figure 4.14. The minutes without inotropes are not included in this distribution, but are present in the dataset. Approximately two-thirds of the windows do not have inotropes.

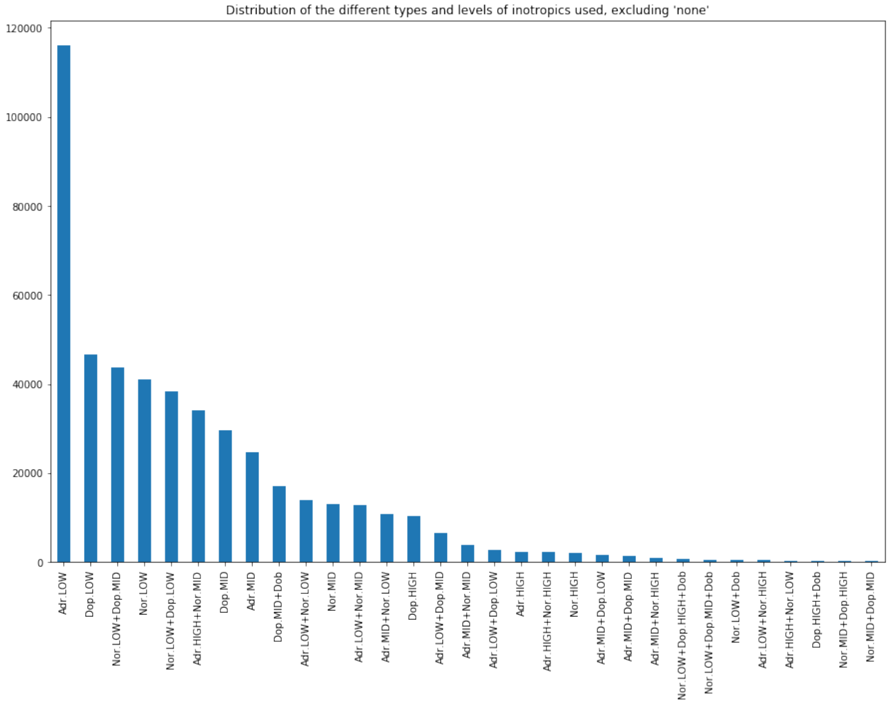


Figure 4.14: Distribution of the interventions in the PICU dataset (excluding *None*)

4.4.2 Interventions OR

The distribution of the types of pump fluid is given in Figure 4.15. Because of interpretability and size, only the most common and meaningful types were selected. The pump fluids selected for the model are marked red. These were manually divided into levels.

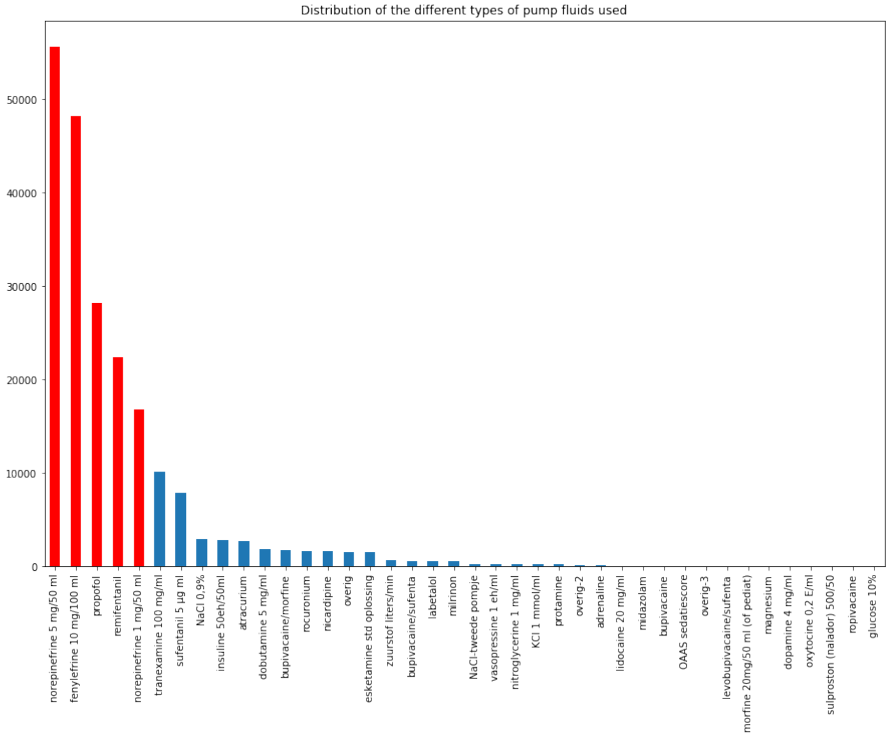


Figure 4.15: Distribution of the interventions in the OR dataset

5

Testing

In order to properly test the performance of this model, it is compared to other models. For this, a set of metrics and models had to be selected.

5.1 Metrics

To assess the performance of the model, three metrics were chosen: the accuracy (both overall and in problematic categories), the Concordance index and the Root Mean Squared Error (RMSE).

5.1.1 Accuracy

The accuracy is the percentage of cases in which the predicted category was the same as the actual next category. While the overall accuracy does give some insight into the performance, the interest is mainly in the 'problematic' categories. For example, the patient's blood pressure might fluctuate within a normal range. This fluctuation is not of interest for this research. By merging these normal categories as mentioned in section 3.1.7, the influence of this fluctuation was minimized in the prediction.

Another possible way to isolate the accuracy of interest is to look at the accuracy per category. The accuracy in a certain category c was defined as the percentage of cases in category $c_{t=i}$ of which the next category $c_{t=i+1}$ was predicted correctly. In other words, the accuracy for category c is the accuracy over all observations with *origin* category c.

This is used in the metrics by defining the accuracy in the problem categories as the sides accuracy. The sides accuracy is defined as the mean accuracy for all categories of which at least one variable is in a problematic range (i.e. in $[ABP0, HF2, SAT0]$).

The difference between the overall accuracy when merged and the sides accuracy is that the overall accuracy merged takes the cases into account in which a patient moves from the center category to a problem category, whereas these are excluded in the sides accuracy. While it might be interesting to predict these changes, they may also consist of sudden declines in the patient that are either unpredictable or result from outside influences that are not represented in the data. For example, in the OR data it is likely that such a decline could be the result of the progress of the surgery.

In terms of accuracy, the model performance is considered good if the accuracy is significantly higher than predicting no change. No change means that the category in the current window is predicted as the category for the next window. As generally in more than half of the cases there is no change between two consecutive windows, predicting no change already provides an accuracy of over 50%.

5.1.2 C-index

The Concordance index or C-index is the area under the Receiver-Operator Curve (ROC) [5]. It measures the goodness of fit in a binary outcome. The ROC-curve plots the true positive rate against the false positive rate. A high C-index shows that the model is good at discriminating between the problematic and normal cases. A

C-index of over 0.7 is considered acceptable and a C-index of over 0.8 indicates a strong model. However, the interpretation of C-index also depends on the incidence.

In order to calculate the C-index for this model, the statistic is split per variable. As the actual and predicted categories are intersections of these variables, this first has to be split. For each variable, the problem category ($AB0$, $HF2$ or $SAT0$) is labeled as 1 and the other two categories are labeled as 0. This is done for the actual categories and the predicted categories and results in two binary vectors per variable.

For example, given the actual category $ABP1 \cap HF2 \cap SAT0$ and the predicted category $ABP0 \cap HF2 \cap SAT0$, the resulting vectors for ABP are $ABP_{actual} = [0]$ and $ABP_{predicted} = [1]$, because $ABP1$ is not the problem category and $ABP0$ is.

These vectors are generated for all predictions for each variable and are then used to calculate the C-index.

5.1.3 Root mean squared error

Finally, the root mean squared error (RMSE) is calculated [2]. The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

where $\hat{\boldsymbol{y}}$ is the prediction and $\boldsymbol{y}$ is the actual observation. The prediction is defined as the prototype (or mean) of the predicted category. The actual observation is taken as the original window. However, in order to avoid skewed results because of outliers in the original data, a median filter of the same width as the SAX method it is compared to is applied. This filter has a width of 5 in the PICU data and a width of 3 in the OR data.

This provides a good measure of the average distance from the actual value to the predicted value. This takes two aspects of the model into account. First, there is a distance caused by categorizing the observations and representing each category by its mean. Secondly, an incorrect category prediction will increase the distance. However, if the original observation is at the edge of the cluster, near the incorrectly predicted cluster, the distance might not increase by a lot.

The RMSE in itself does not provide an indication of the model's performance, but can be used to compare types of models and to see if the average error is acceptable in a clinical perspective.

The RMSE is only calculated in the unmerged model, as merging the center groups multiple states of multiple variables. This inherently increases the error and changes the prototype, but only in categories that are of less interest. It is therefore not a useful performance measure. As the RMSE is mainly included to evaluate the influence of adding the SAX step to the model, it is sufficient to calculate it for unmerged cases only.

5.2 Comparisons

These metrics give an indication of the performance of the model in itself, but can also be used compare the model to other models or variations of the current model. The base model is the model described in Chapter 3 using K-Means and SAX to categorize the data and combining the current, next and potential previous categories with the interventions to generate a probability matrix. There are three models that are interesting to compare the performance of this model to: the no change model, the model with only K-Means and no SAX and the model with K-Means and SAX without interventions.

5.2.1 No Change model

As mentioned in section 5.1.1, the accuracy of a model is compared to the accuracy of simply predicting the current category. As there are many situations in which the category does not change between two windows, it is important to compare this. By predicting the same category as the current window, the model could have a reasonable accuracy without having any predictive value. It is therefore important to see whether the current model performs better than the No Change model.

5.2.2 K-Means

In order to check the usefulness of including the SAX step in the data preparation stage, the model is checked against a model without this symbolisation step. In order to avoid the outliers and reduce the dimensionality of the data, a simple median filter is applied instead. This median filter uses the same width as the segments in the SAX model it is compared to. This width is 5 minutes in the PICU dataset and 3 minutes in the OR dataset.

The expectation is that the RMSE is lower in the K-Means only model than in the model with SAX if the accuracy is similar, as K-Means directly tries to minimize the distances within each cluster. While K-Means might get stuck in a local optimum, this effect should be minimized by taking the mean of multiple iterations.

It is expected that the abstraction caused by applying SAX increases this error. It is therefore interesting to see whether this increase is large enough to be problematic.

As mentioned in section 2.2, it is expected that using K-Means without SAX causes the center category to change with regard to the problematic categories. In order to test this, the number of time windows in the side is recorded as well.

5.2.3 Model without interventions

In order to quantify the influence of the interventions, the model is also run without interventions. The model remains unchanged, except that the intervention is set to *none* for each entry in the prepared dataset.

As the end-tidal vapor is (the direct result of) an intervention, it is also excluded from this model.

The difference in the performance caused by excluding the interventions can provide insight into the extra information given by the intervention when the patient status is known. This insight is necessary to answer the question whether the influence of the interventions can be isolated.

5.3 Test method

After the data preparation step is performed, the training and testing of the model is run using 10-fold cross validation. This means that the dataset $\mathcal{D}$ is randomly divided into 10 disjoint subsets $s_i \in \mathcal{D}$ such that $\bigcup_{i=1}^{10} s_i = \mathcal{D}$. For each of the 10 iterations or folds i, subset s_i is taken as the test set, while $\mathcal{D} \setminus s_i$ is used as the training set. The training set is passed to the training method to produce a probability matrix. The test method uses the test set to predict each category and compare it to the actual value. For each row in the test set, the test method passes the current category, intervention and potential previous categories to the prediction method, combined with the probability matrix. The returned prediction can then be compared to the actual category to calculate the performance of the model.

The overall accuracy and accuracy per category are calculated per fold and the mean is returned. All predictions are returned combined with the true category, in order to calculate the C-index and the RMSE.

In order to get a more robust result and reduce the influence of randomness in the data preparation step, the combination of preparation and 10-fold cross validation is repeated 10 times and the mean is taken for each metric. In Chapter 6, these mean values are given as results.

5.4 Parameters

The parameters chosen for these tests are as follows:

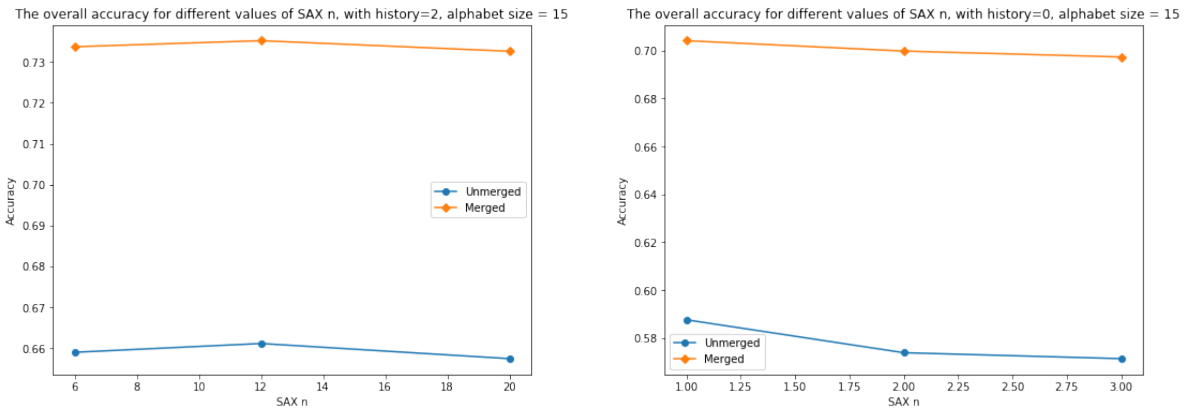
	PICU	OR
SAX $n_{segments}$	12	2
SAX *alphabet_size*	15	15
History h	2	0
Number of clusters k	3	3

The parameters were based on a parameter sweep. The options for the number of segments were chosen such that the window size is divisible by it. As can be seen in Figure 5.1a, the model performs best on the PICU dataset for $n = 12$. In the OR dataset, the model performs best with $n = 1$. However, as this would remove all variation from the window, $n = 2$ was chosen. This performs only slightly worse.

The alphabet size was set to 15. As can be seen in Figure 5.2a, this yields the best performance for the PICU dataset. In the OR dataset, an alphabet size of 20, performs slightly better in some iterations. As this difference is small and did not occur consistently, an alphabet size of 15 was chosen.

The number of clusters k was set to three. A higher number of clusters caused an exponential increase in the number of prediction categories, and therefore worsened the performance and caused problems with the interpretability.

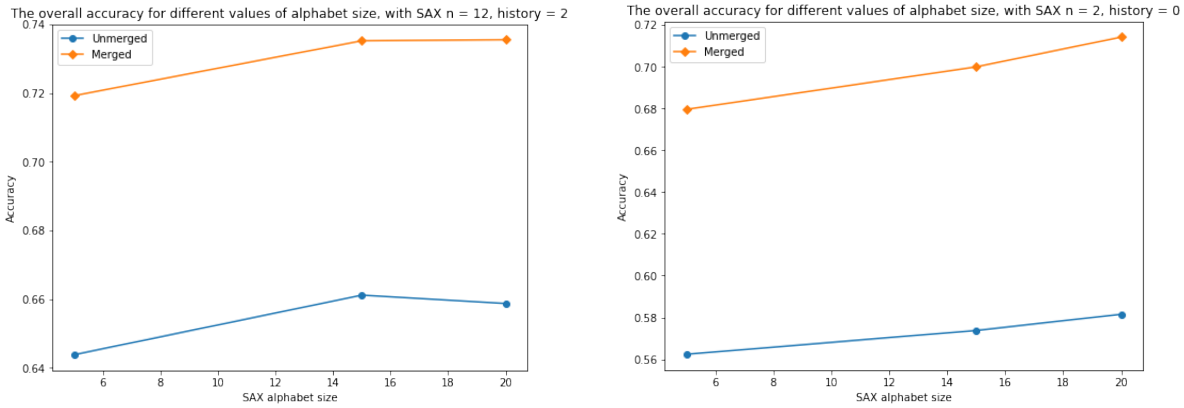
Adding history has a different effect in the two use cases. In the PICU dataset, adding two hours of history showed significant improvement in the accuracy of the model. This improvement can be seen in Figure 5.3a. In the OR dataset, adding any window as history only worsened the performance, as shown in Figure 5.3b. This sweep was only performed in the unmerged model, as this trend was clear and the time and space requirements of running the OR model with history were very large.



(a) SAX number of segments in the PICU dataset

(b) SAX number of segments in the OR dataset

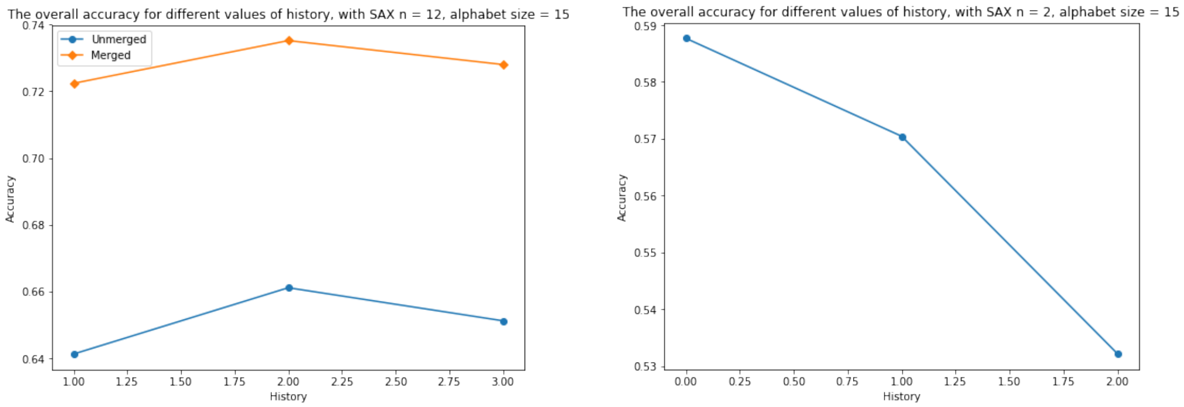
Figure 5.1: Parameter sweep for the number of segments



(a) SAX alphabet size in the PICU dataset

(b) SAX alphabet size in the OR dataset

Figure 5.2: Parameter sweep for the alphabet size



(a) History in the PICU dataset

(b) History in the OR dataset

Figure 5.3: Parameter sweep for the history

6 Results

For each metric, the models are compared and a comparison is made between the PICU and OR datasets. All reported values are the mean over the iterations and over the 10-fold cross-validation.

For all metrics except the RMSE, the values are given both for the merged and the unmerged model. In the unmerged model, all time window categories are the intersection of the categories of their variables. In the merged model, a time window category that does not contain any problematic category ($ABP0, HF2orSAT0$), is represented as $Center$. In the current and previous window, the $Center$ category can be intersected with covariables.

6.1 Accuracy

6.1.1 PICU

The accuracy for the PICU dataset can be found in Figure 6.1. As can be seen, the No Change model performs worse than the other models. The difference is significantly larger in the sides ($\Delta = 0.12$) than in the edges ($\Delta = 0.052$). For each model, the overall accuracy improves significantly when the center categories are merged ($\Delta = 0.066$), but the sides accuracy barely changes ($\Delta = 0.0034$).

The accuracy is slightly lower when the interventions are removed, but the change is very small. The mean decrease in accuracy between a model and its counterpart without interventions is 0.013 in the overall accuracy and 0.011 in the side accuracy.

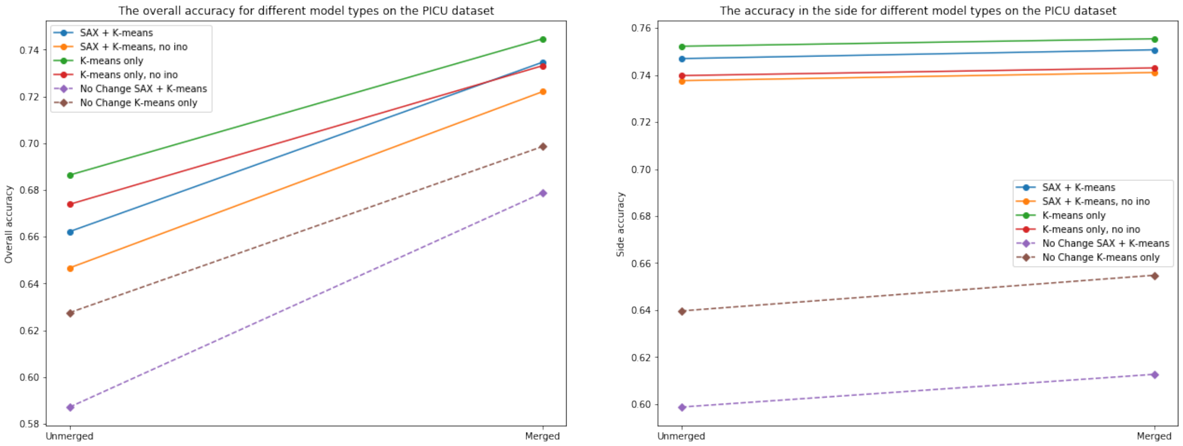
The accuracy is slightly higher when the SAX step is removed from the model, but the change is again very small. The mean increase in accuracy between each model and the counterpart without SAX is 0.018 in the overall accuracy and 0.0035 in the side accuracy.

6.1.2 OR

The accuracy for the OR dataset can be found in Figure 6.2. As can be seen, the No Change model performs significantly worse ($\Delta = 0.16$) than the other models for the side accuracy. In the overall accuracy, the No Change accuracy is very close to the other models and even occasionally performs better. For each model, the overall accuracy improves significantly when the center categories are merged ($\Delta = 0.13$), but the sides accuracy barely changes ($\Delta = 0.010$).

The overall accuracy is slightly higher when the interventions are removed, but the change is very small. The mean increase in overall accuracy between a model and its counterpart without interventions is 0.013. The side accuracy is lower when the interventions are removed, with a larger difference ($\Delta = 0.037$).

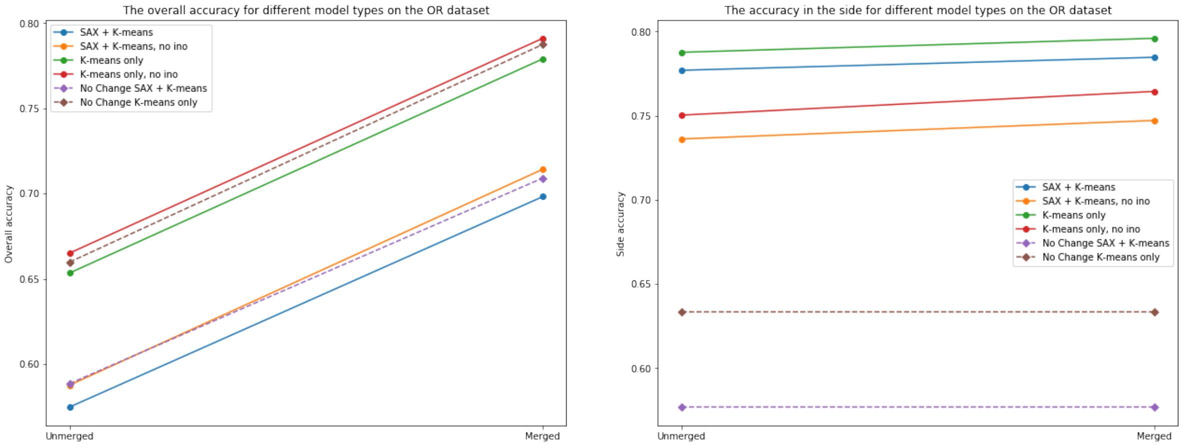
The accuracy is higher when the SAX step is removed from the model. The mean increase in accuracy between each model and the counterpart without SAX is 0.078 in the overall accuracy and 0.013 in the side accuracy.



(a) Overall accuracy in the PICU dataset

(b) Side accuracy in the PICU dataset

Figure 6.1: The accuracy for the PICU dataset



(a) Overall accuracy in the OR dataset

(b) Side accuracy in the OR dataset

Figure 6.2: The accuracy for the OR dataset

6.1.3 Comparison

As can be seen in Table 6.1, the model has a higher overall accuracy on the PICU dataset than on the OR dataset. However, the side accuracy is higher for the OR dataset.

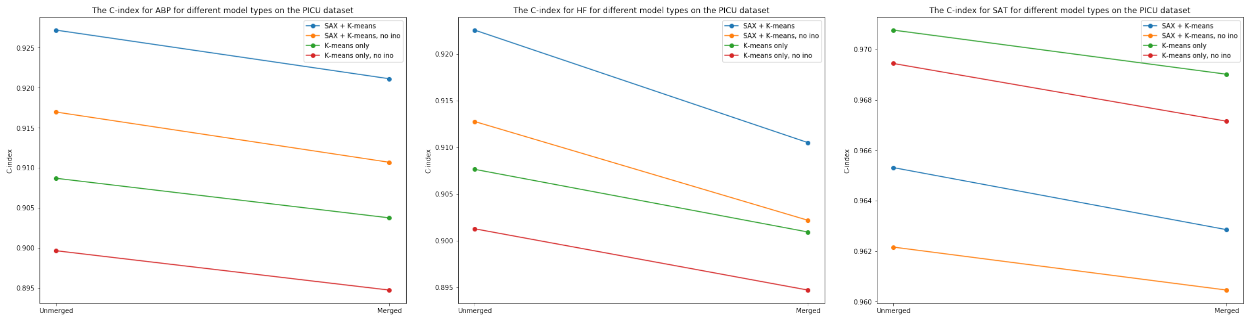
Variable	PICU	OR
Overall accuracy	66.2%	57.5%
Overall accuracy merged	73.4%	69.8%
Side accuracy	74.7%	77.7%
Side accuracy merged	75.1%	78.5%

Table 6.1: Comparison of the accuracy for both datasets for the SAX + K-Means model (with interventions)

6.2 C-index

6.2.1 PICU

The values for the C-index for the different models for the PICU dataset can be found in Figure 6.3. Merging the center categories or excluding the interventions slightly decreases the C-index. All values are above or around 0.9. For ABP and HF, the highest C-index occurs for the SAX and K-Means model. For SAT, the C-index is slightly higher for the K-Means only model.



(a) C-index for ABP

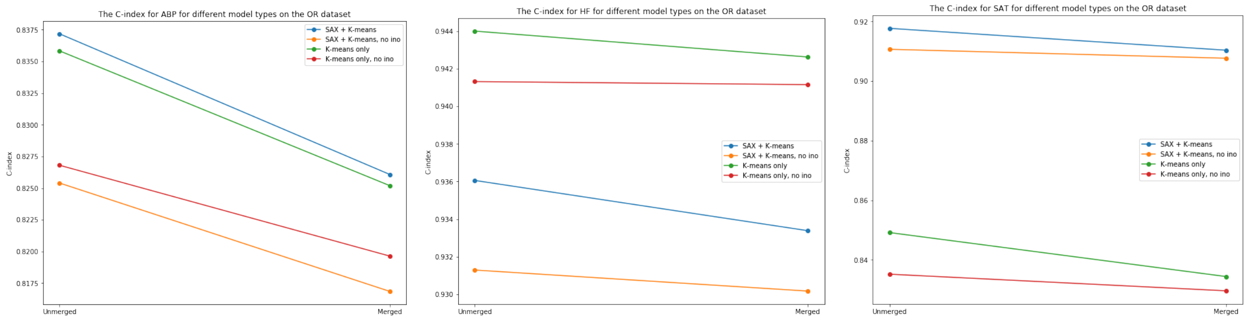
(b) C-index for HF

(c) C-index for SAT

Figure 6.3: The C-index in the PICU dataset

6.2.2 OR

The values for the C-index for the different models for the OR dataset can be found in Figure 6.4. Merging the center categories or excluding the interventions slightly decreases the C-index. All values are between 0.8 and 0.95. For ABP and SAT, the highest C-index occurs for the SAX and K-Means model. For HF, the C-index is slightly higher for the K-Means only model.



(a) C-index for ABP

(b) C-index for HF

(c) C-index for SAT

Figure 6.4: The C-index in the OR dataset

6.2.3 Comparison

As can be seen in Table 6.2, the C-index differs a lot per variable. For ABP and SAT, the C-index is significantly higher on the PICU dataset than on the OR dataset. For HF, the values are quite close, with the model performing slightly better on the OR dataset.

Variable	PICU	OR
ABP	0.927	0.837
HF	0.923	0.936
SAT	0.965	0.918

Table 6.2: Comparison of the C-index for both datasets for the SAX + K-Means model (unmerged, with interventions)

6.3 RMSE

6.3.1 PICU

The RMSE for the different models for the PICU dataset can be found in Table 6.3. In all three variables, the RMSE is lower for the K-Means only model than for the SAX + K-Means model and lower for the model with interventions than the accompanying model without interventions. There is very little difference in RMSE for ABP between the models, while there is a larger difference between the models in the RMSE for the variables HF and SAT.

Model	ABP (mmHg)	HF (bpm)	SAT (%.)
SAX + K-Means	11.29	14.91	4.44
SAX + K-Means without interventions	11.35	15.02	4.47
K-Means only	11.10	14.20	3.76
K-Means only without interventions	11.16	14.30	3.78

Table 6.3: The RMSE in the PICU dataset

6.3.2 OR

The RMSE for the different models for the OR dataset can be found in Table 6.4. For HF and SAT, the RMSE is lower for the K-Means only model than for the SAX + K-Means model and higher for the model with interventions than the accompanying model without interventions. For ABP, the SAX + K-Means model has a slightly lower RMSE than the K-Means only model, and the model with interventions has a lower RMSE than the accompanying model without interventions.

Model	ABP (mmHg)	HF (bpm)	SAT (%.)
SAX + K-Means	12.68	28.55	1.69
SAX + K-Means without interventions	12.88	28.07	1.70
K-Means only	12.73	19.97	1.58
K-Means only without interventions	13.00	19.81	1.60

Table 6.4: The RMSE in the OR dataset

6.3.3 Comparison

As can be seen in Table 6.5, there is a large difference between the RMSE of the model on both datasets. While the RMSE for ABP is similar between the two datasets, this is not the case for HF and SAT. The model has a significantly higher error for HF on the OR dataset than on the PICU dataset. The opposite happens for SAT, where the model has a significantly lower error on the OR dataset.

Variable	PICU	OR
ABP	11.3 mmHg	12.7 mmHg
HF	14.9 bpm	28.6 bpm
SAT	4.44 %.	1.69 %.

Table 6.5: Comparison of the RMSE for both datasets for the SAX + K-Means model (unmerged, with interventions)

6.4 Cluster centers and number in side

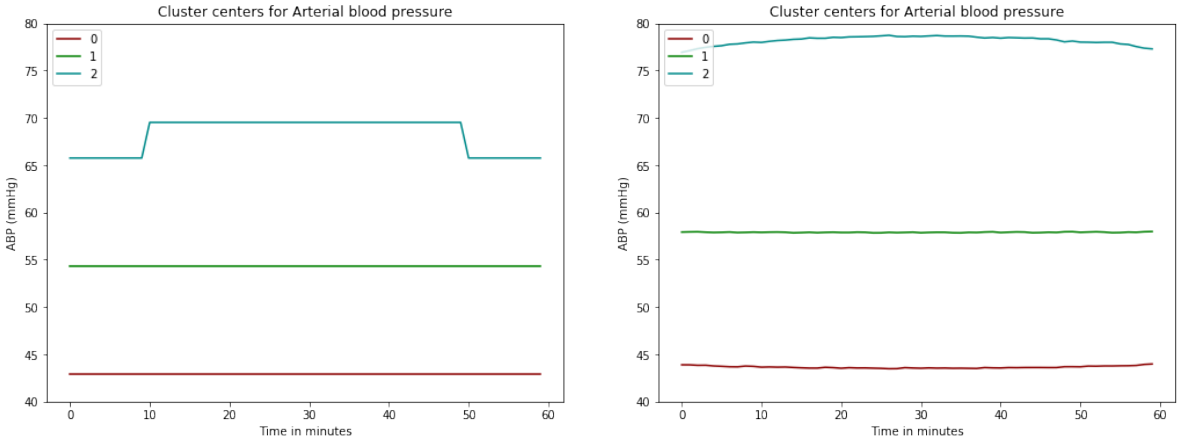
In order to see the influence of SAX on the cluster division, the prototypes or cluster centers are extracted. As the cluster centers are in terms of SAX symbols, these are then transformed back to the unit of the variable. As the SAX symbols are applied using breakpoints, the reverse transformation replaces a SAX symbol by the center of that range. For example, a symbol associated with ABP values of $50 \leq 60$mmHg is transformed to 55mmHg. As the SAX transformation also reduced the dimensionality, these reverse transformed values are then repeated for the length of the SAX segment (i.e. 5 minutes in the PICU dataset and 3 minutes in the OR dataset).

In Figures 6.5 to 6.7, the cluster centers for the haemodynamic parameters in the PICU dataset can be found. The figures show that the cluster centers change when SAX is not used. In the model with only K-Means, the outer cluster centers move further outward. The size of this move differs per variable.

This shift can cause the problematic category ($ABP0$, $HF2$ or $SAT0$) to change with regard to the PEWS score. For example, the clustering of SAT by the SAX + K-Means model corresponds directly to the PEWS score. The problematic category $SAT0$ gets a score of 2, which is the highest score for that variable and corresponds to an oxygen saturation under 91%. In the K-Means only clustering, a saturation of less than 91% is between the middle and high cluster prototype. Therefore, this clustering is not easily interpretable in terms of the PEWS score, as part of the middle cluster is already in a problematic state.

In the ABP and HF cluster centers, this change is not as large.

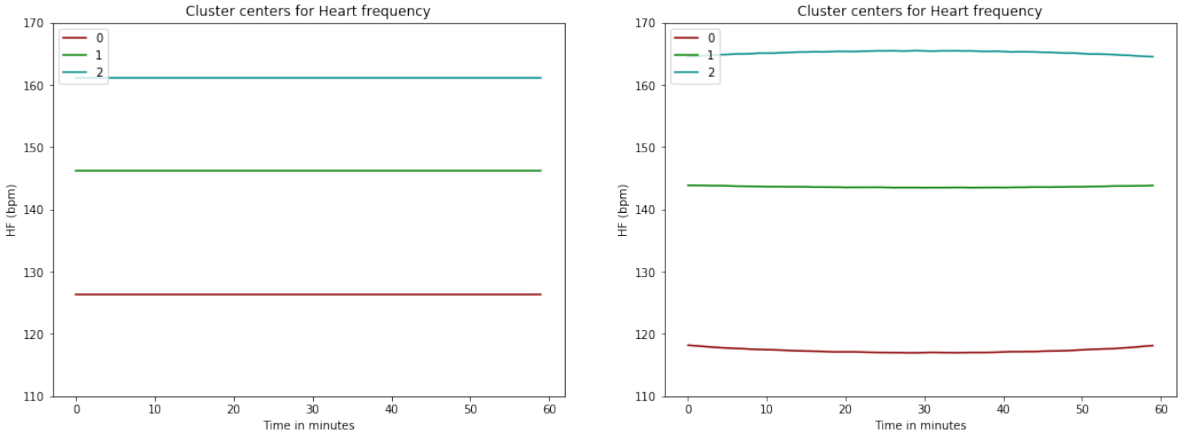
In Figures 6.8 to 6.10, the cluster centers for the haemodynamic parameters in the OR dataset can be found. The same shift can be seen here. A large difference can be seen in the clustering of HF in Figure 6.9. Here, the cluster center problematic category HF2 moves up significantly. While the prototype for HF2 in the SAX + K-Means model corresponds with a EWS score of 1 or higher, the corresponding prototype in the K-Means only model only includes values with a score of 3. The same happens for SAT0 in Figure 6.10.



(a) Cluster centers for ABP for the SAX + K-Means model

(b) Cluster centers for ABP for the K-Means only model

Figure 6.5: The cluster centers for ABP in the PICU dataset



(a) Cluster centers for HF for the SAX + K-Means model

(b) Cluster centers for HF for the K-Means only model

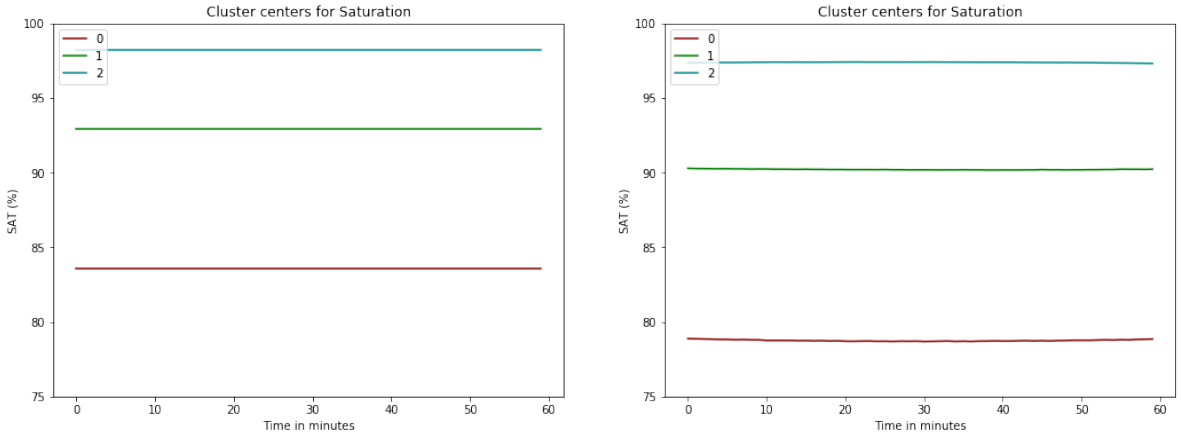
Figure 6.6: The cluster centers for HF in the PICU dataset

6.4.1 Number of time windows in side

These changes in the clustering can also be seen in the distribution of the number of time windows between the side and the center. This distribution can be found in Table 6.6. The table shows that less time windows are in any problematic (side) category in the K-Means only model than the SAX + K-Means model. This decrease is small in the PICU dataset, but significantly larger in the OR dataset.

Variable	PICU	OR
Total number of windows	75567	456053
SAX + K-Means	50453 (66.8%)	236004 (51.7%)
K-Means only	49923 (66.1%)	168025 (36.8%)

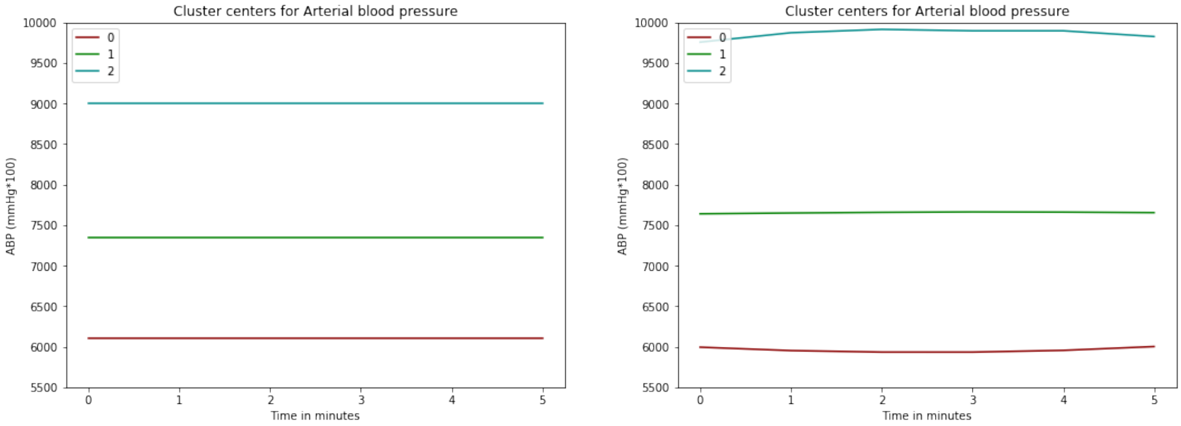
Table 6.6: Comparison of the mean number in the side for both datasets for the SAX + K-Means model versus the K-Means only model



(a) Cluster centers for SAT for the SAX + K-Means model

(b) Cluster centers for SAT for the K-Means only model

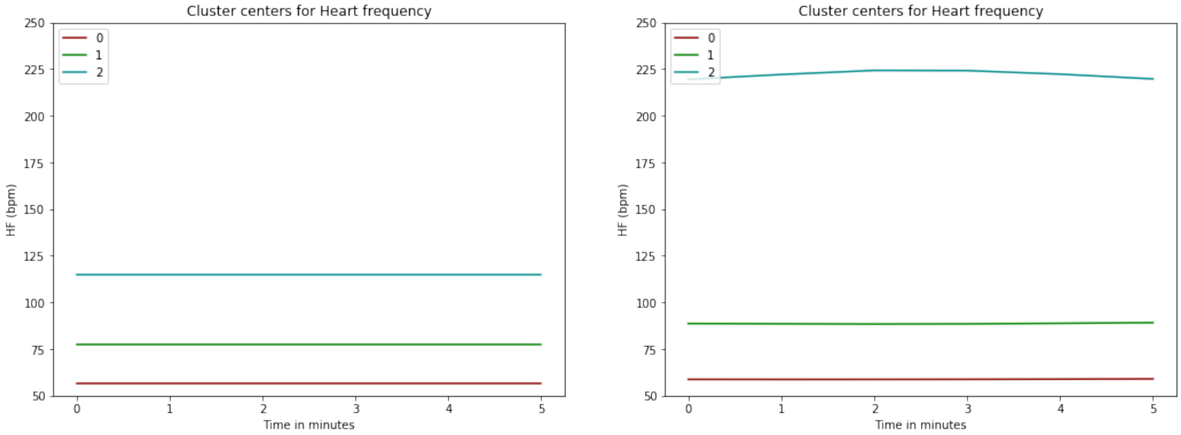
Figure 6.7: The cluster centers for SAT in the PICU dataset



(a) Cluster centers for ABP for the SAX + K-Means model

(b) Cluster centers for ABP for the K-Means only model

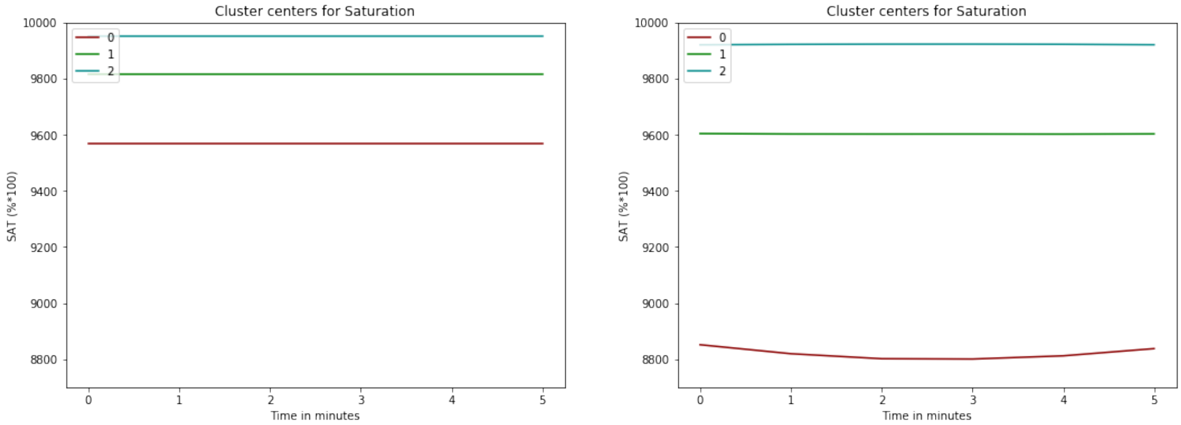
Figure 6.8: The cluster centers for ABP in the OR dataset



(a) Cluster centers for HF for the SAX + K-Means model

(b) Cluster centers for HF for the K-Means only model

Figure 6.9: The cluster centers for HF in the OR dataset



(a) Cluster centers for SAT for the SAX + K-Means model

(b) Cluster centers for SAT for the K-Means only model

Figure 6.10: The cluster centers for SAT in the OR dataset

7 Discussion

7.1 Performance

As seen in the previous chapter, the model performs sufficiently well. In all cases, the accuracy of the K-Means+SAX model is significantly higher than the No Change model for the problematic (side) categories. With a side accuracy of 75.1% to 78.5%, the results are promising.

The C-index ranges from 0.84 to 0.97 and is therefore significantly above the threshold of 0.7. As the percentage of the time windows in the problematic category for each variable is generally in the range of 20% to 35% of the total number of time windows, this can be considered good performance. This C-index is significantly higher than to be expected from the accuracy. This is the case because the accuracy per variable is significantly higher than the overall accuracy. Within each variable, the category remains unchanged more often than the overall category. The accuracy of the No Change model per variable is generally above 80% for both datasets, with each other model improving on this. There is one exception to this range. The No Change accuracy for ABP in the OR dataset is only 72.5%. This corresponds with the lowest C-index as found in Table 6.2.

There is some difference between the PICU data and the OR data. The higher overall accuracies and C-index for the PICU dataset may be explained by the age differences within the dataset. While filtering the dataset to only include data from patients of less than one year old does reduce the influence of age, there is still some variation in the normal ranges. This might influence the categorization, as the age is implied in the data and therefore might change the likelihood of certain category changes. This variation may cause less movement from one category to another within each variable.

Another possible explanation is the higher degree of outside influence in the OR dataset. In the OR data, a surgery is taking place. Changes within the 'normal' ranges of the variables could therefore be more difficult to predict from the available data, as no information about the progress of the surgery is used. In the problematic categories, changes in the haemodynamic status could be more related to the intervention as action is taken to return to a normal range, which might explain the higher performance in these categories. The difference in RMSE between the datasets is caused by the difference in the distribution of the values and thereby the distribution of the cluster centers.

There are also some differences between the performance of the SAX + K-Means model and the K-Means only model. As expected, the RMSE is slightly higher in the SAX + K-Means model than in the K-Means only model, but the difference is small. The accuracy is very similar, but slightly higher for the K-Means only model. This can be explained by the larger center category, as seen in Table 6.6.

In conclusion, the model performs sufficiently. While adding SAX does not increase the performance of the model, it does change the category division. As seen in section 6.4, the cluster centers change significantly when the SAX step is excluded. The center category becomes larger and the cluster center corresponding to the problematic category becomes more extreme. In the case where less extreme category definitions are desirable, and the focus is on the problematic categories, the model with SAX can therefore be considered the more suitable model.

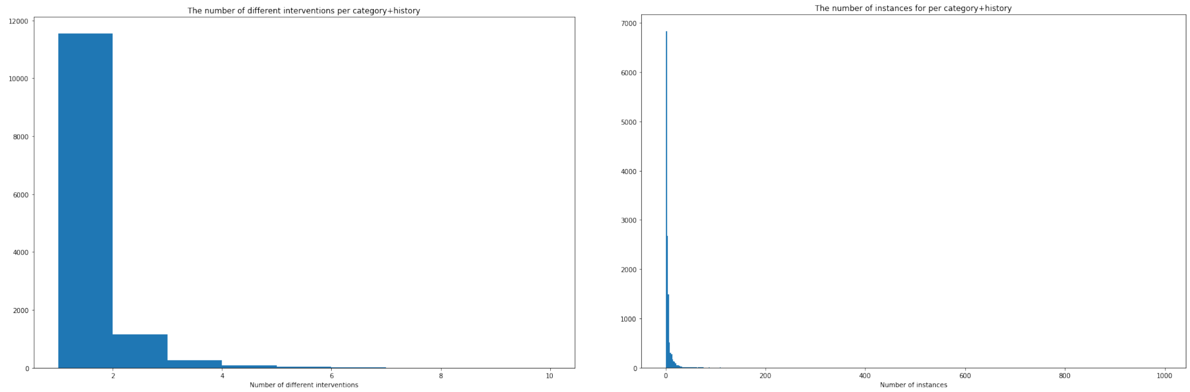
7.2 The role of intervention

As can be seen in Figure 6.1, there is very little change in the accuracy when the interventions are excluded. This suggests that the correlation between the patient status and the medication they receive is very high.

In order to confirm this, the distribution of the interventions needs to be examined further. In Figure 7.1a, the number of unique interventions per combination of current category and history ($h = 2$) is shown for the (unmerged) PICU data. It shows that in the majority of the cases, there is only one type of intervention present for each combination. The mean number of interventions is 1.17, with the median at 1. The high level of detail in the categorization of the situation achieved by adding history and covariables causes a decrease in the number of datapoints in each category and therefore a decrease in variation within the categories. The number of time windows per category (and history) can be seen in Figure 7.1b. The mean number of time windows per category is 5.77, with the median as low as 2. This shows that next to the small number of intervention types, there is also a small total number of time windows left in each cell of the probability matrix.

Merging the center increases the number of time windows per category and the variation, but only slightly. Cases with a single type of intervention still hold a large majority.

Therefore, the high level of detail created by including covariables and history causes the number of time windows in each cell to be very small and generally associated with only one type of intervention. This means that in the PICU dataset, the small difference in performance by excluding the interventions is caused by a high correlation between the patient status and the intervention and a small number of datapoints for each status.



(a) A histogram of the number of unique types of intervention for each combination of current category and history ($h = 2$) for the PICU data (unmerged)

(b) A histogram of the number instances for each combination of current category and history ($h = 2$) for the PICU data (unmerged)

Figure 7.1: Distribution of instances and types of intervention for the PICU dataset

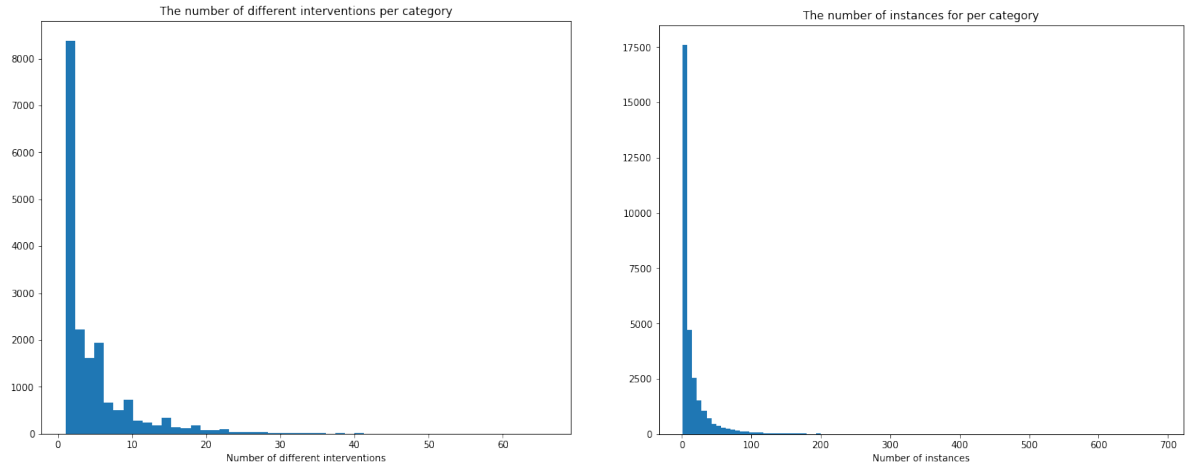
As there are significantly more data points in the OR dataset and no history is included there, there should be more datapoints in each category. This can be seen in Figure 7.2b. The median number of instances per combination of category and patient information is 6 (mean of 15.5), whereas the median number of instances in the PICU dataset was only 2 (mean of 5.77). There are also significantly more combinations of interventions, as there are five types of medication used with five levels each. When vapor (with three categories) is included as an intervention, there are over 1000 unique combinations of interventions. The PICU dataset contains only 32 possible combinations of interventions.

This results in a slightly larger number of different interventions for each combination of status and patient information. As can be seen in Figure 7.2a, this variation is larger than in the PICU dataset, with a mean of 4.8 (and a median of 3) unique intervention types per cell in the probability matrix. While it is less clear than in the PICU dataset, this still suggests a high level of correlation as the number of unique types of interventions per cell is very small compared to the total number of interventions.

7.3 Conclusions

In Chapter 1, two connected research questions were defined. The first question was whether the change in the status could be predicted accurately. As explained in section 7.1, the performance meets the criteria set in Chapter 5. The model significantly outperforms the No Change model in the categories of interest (side accuracy) and the error increases only a little by including the SAX step.

The second question was whether the role of the intervention could be isolated. As explained in section 7.2, this is not possible as the intervention is too highly correlated with the patient's status.



(a) A histogram of the number of unique types of intervention for each current category for the OR data (unmerged)

(b) A histogram of the number instances for each current category for the OR data (unmerged)

Figure 7.2: Distribution of instances and types of intervention for the OR dataset

7.4 Practical use in medicine

As shown above, the accuracy and error of the model seem promising. This shows that the method could potentially have practical use. The model is able to predict any situation that has been encountered before.

It is not possible to isolate the influence of the interventions on the prediction. It seems that any prediction made by the model implicitly includes the intervention in the patient's status. This problem might be solved by including the intervention explicitly in the model, even though removing it makes little difference in the performance. When it is included, the model can still predict the situation if the actual intervention corresponds with the intervention present in the same situation in the training set. If so, a prediction can be given. If not, it might not be possible to predict the next status.

This generalizes to the problem that only previously seen situations can be predicted by this model. Combinations of categories and history that do not occur in the current dataset cannot be predicted. This is only problematic if there are many such cases. If most new situations fall into one of the currently filled cells in the probability matrix, there is no problem. By making the model as specific as it is now, it performs best on the current datasets. However, when new data is introduced, the current level of detail might cause it to fall in an unseen category. Therefore, the level of detail might cause the model to be overfit to the current dataset and thereby be weak in predicting new, actual cases.

Another problem is that a trend from a problematic status towards a normal status might be important even if it is not large enough for the status to move to a normal category. In order to combat this, finer-grained categories might be necessary for the model to be useful in practice. However, this problem is inherent to any model that uses categorization. Any clustering will cause the loss of some information, unless the number of clusters becomes prohibitively large. An increase in clusters would also make the model even more specific, thereby worsening the problem of potentially overfitting and losing compatibility to new cases.

Therefore, using this model to predict a patient's future status might help provide doctors with a quick, general idea of the severity of the situation over time. The performance of the model seems acceptable for practical application, but further examination would be needed to evaluate whether the level of detail in the categories is sufficient. It would also be necessary to explore whether new situations correspond to the available combinations of current (and previous) categories and interventions in the probability matrix and therefore whether they could be predicted.

7.5 Further research

This research has brought many new questions that need to be researched further. First of all, a lot of datapoints are discarded because only the windows with no missing data are used. This could potentially be improved by defining a maximum number of (consecutive) missing values. A problem with these missing values is that they are not necessarily missing at random.

As mentioned in section 7.1, the performance of model on the PICU dataset is influenced by the difference in age. Even though the dataset has been filtered to only include patients of less than one year old, there is still a noticeable change in the reference values with regard to age. This might cause the model to partially categorize patient time windows by age instead of by severity of the situation. This might influence the likelihood of moving between two specific categories and make the interpretation of the results more difficult. A possible solution to this is to take the difference from the baseline as the status, with an age-dependent baseline. This can be done in multiple ways. For example, the direct deviation with regard to the baseline can be taken (e.g. -5 mmHg). The problem with this might be that the absolute deviation may have different meaning depending on the baseline. Instead, the difference could be measured in terms of standard deviations. However, this introduces the problem that the standard deviation would have to be calculated by age group, which requires a cutoff between groups and might therefore be problematic with a baseline that is a continuous function of the age. Therefore, this change requires further research and was outside of the scope of this research.

In section 7.4, it was shown that a potential problem with this model is that it can only predict cases that have been seen before. Categorizing generalizes the situation somewhat, but the addition of intervention, covariables and history makes the situation more specific. This means that there are many situations that might occur in practice, but cannot be predicted because the corresponding cell in the probability matrix has not been filled with previous data. While generalizing the model (e.g. by removing history) might solve this, it would also significantly decrease the performance. It might therefore be useful to extend the model such that an unseen case is generalized to a combination of similar cases in order to allow for prediction. However, it would need to be researched whether such an extension adds any predictive performance, as it would need to be accurate enough to actually benefit the model.

The third and potentially most important question lies in the role of the categories with regard to the time. In the current model, the category of the current time window w_i is taken and the majority next category in its cluster for w_{i+1} is returned. This means that the status ‘jumps’ from one category to another. Another possible approach would be to take the mean or median of the measurements in the next window w_{i+1} for all members of the cluster found in w_i. If all of the members of the current cluster of w_i are in the same cluster at w_{i+1} and there are no other members of the cluster at w_{i+1}, this gives the exact same result as the current model. However, this will generally not be the case. Not only can there be variation within a cluster such that the majority category for the next window does not contain all of the current members, but there can also be new members in the cluster of the next window that skew its prototype.

Comparing the current model to a model that returns the mean or median of all members for the next time window (instead of the prototype of the majority next category) can provide more insight into the quality of the categorization and the time factor in the prediction. The time factor differs as the current model assumes that patient windows that are currently in a similar state may differ greatly in the next point in time. Taking the mean or median assumes that patient windows currently grouped together stay together for the next hour. Whether this differs greatly will depend on the nature of the time relations in the dataset. It also depends on the variation and level of detail in the model. As shown in section 7.2, there is little variation left within each cell of the probability matrix. When there is no variation, the difference in the time component is difficult to extract as often 100% of the datapoints in a cell are in the same next category. There may however be some influence of other datapoints that occur in the next category and may skew the prototype.

The comparison may also give some insight into the quality of the categorization. If there are distinct categories, a model that chooses the majority category should perform better, as it would have a small distance to the actual value in the majority of the cases. If there are no distinct categories, the cutoff between adjacent categories might be arbitrary and a mean or median might perform better.

As this small change to the model could provide great insight in the nature of the relations in the dataset, it is an interesting direction of further research.

It might also be useful to further explore the definition of interventions. In the current model, the intervention was taken as the current status of the continuous medication. In other words, the time windows were defined independent of the interventions and the intervention used was based on the window. Another possible approach might be to take any change in the medication and take the time window before that to predict the window afterwards. This would define the intervention as the direct change of the status of the medication. Defining an intervention as such might provide a different insight in the influence of the interventions as there is a more direct change. It would therefore define a more event-based model.

This approach would ignore all medication that is the same as the previous minute, as this would not be considered an intervention. However, this means that a deliberate choice to keep the medication at the current state would also be ignored, even though it might add valuable information to the model.

This suggests that the difference between the current model and the event-based model lies in the interpretation of the medication status. When the intervention is defined as any deliberate choice made with regard to the medication, neither model provides a realistic interpretation of the situation. The current model assumes that the medication at each minute is a deliberate choice, whether it was changed or not. The event-based model would assume that only changes are deliberate choices, and continuation of the current medication would not be seen as a decision. As neither model corresponds directly to the real situation and the difference is not clear from the data, it might be useful to perform further research into which definition of intervention would be valid.

This could be useful to research as the difference in definition might have a large influence on the ability to isolate the intervention and to draw conclusions on the relations between the intervention and the status.

Finally, the influence of the change in status within a time window is minimized in this model. As could be seen in section 6.4, there is little change over time within each cluster prototype. It seems that all variation within the window has been averaged out. This means that the change in status is only used in the prediction when history is included and only if the trend involves a change in category. Any trend within a category is not used.

Recently, a new extension to SAX has been proposed that does take this trend into account [18]. This new technique SAX_CP extends SAX by detecting trend changes using variable segment sizes to improve the approximation of the data. It might therefore be interesting to see what influence this extension has on the current model.

While the model performs sufficiently well, these are interesting areas of further research that might improve the performance and provide more insight in the relations found in the data.

Bibliography

[1] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[2] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[3] Kaushik Chakrabarti, Eamonn Keogh, Sharad Mehrotra, and Michael Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems (TODS)*, 27(2):188–228, 2002.

[4] Daniel DeMers and Daliah Wachs. Physiology, mean arterial pressure. *Physiology, Mean Arterial Pressure*, Feb 2019.

[5] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[6] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[7] Vijaylakshmi Kamat. Pulse oximetry. *Indian Journal of Anaesthesia*, 46(4):261–8, 2002.

[8] Jordan King. Physiology, cardiac output. *StatPearls*, May 2019.

[9] Chiara Lazzeri, Serafina Valente, Marco Chiostri, and Gian Franco Gensini. Clinical significance of lactate in acute cardiac patients, Aug 2015.

[10] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2):107–144, 2007.

[11] Marco Metra, Gad Cotter, Mihai Gheorghiade, Livio Dei Cas, and Adriaan A. Voors. The role of the kidney in heart failure. *European Heart Journal*, 33(17):2135–2142, 08 2012.

[12] RWD Nickalls and WW Mapleson. Age-related iso-mac charts for isoflurane, sevoflurane and desflurane in man. *British journal of anaesthesia*, 91(2):170–174, 2003.

[13] ASA House of Delegates. asa physical status classification system. *American Society of Anesthesiologists (ASA)*, 2014.

[14] William D Owens, James A Felts, and Jr EL Spitznagel. Asa physical status classifications: a study of consistency of ratings. *Anesthesiology*, 49(4):239–243, 1978.

[15] Chotirat Ann Ralanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, and Gautam Das. Mining time series data. In *Data mining and knowledge discovery handbook*, pages 1069–1103. Springer, 2005.

[16] Voedingscentrum. Bmi berekenen. *Stichting Voedingscentrum Nederland.*

[17] Nederlandse Vereniging voor Klinische Chemie en Laboratoriumgeneeskunde. Asat. *Nederlandse Vereniging voor Klinische Chemie en Laboratoriumgeneeskunde*, Jan 2011.

[18] Hamdi Yahyaoui and Reem Al-Daihani. A novel trend based sax reduction technique for time series. *Expert Systems with Applications*, 130:113–123, 2019.

[19] Guoqiang Zhang, B Eddy Patuwo, and Michael Y Hu. Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 14(1):35–62, 1998.