Identification of Phenotypes in Cardiac Arrest Patient Cohorts

Rebecca lütz

Master's thesis
2021:E44

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Populärvetenskapligt sammanfattning

När hjärtat slutar slå och inte pumpar blod längre, kallas det för hjärtstopp. Det finns olika orsaker som leder till hjärtstopp så som underliggande hjärtsjukdomar eller ventrikelflimmer. Utan behandling dör patienter som har drabbats av hjärtstopp och även med behandling leder hjärtstopp ofta till död eller neurologiska problem.

Målet med detta arbete är att dela in hjärtstoppspatienter i olika grupper där alla patienter inom samma grupp ska vara så homogena som möjligt medan patienter från olika grupper ska vara så heterogena som möjligt. Detta med avseende på de kliniska variablerna som används för att hitta grupper men även med avseende på mortalitet och neurologiskt utfall.

I detta arbete har två olika dataset använts där båda innehåller en blandning av kontinuerliga och kategoriska variabler. En variabel som finns med i båda är den så kallade ”targeted temperature management” variabeln som förkortas TTM och indikerar om en patient har kylts ned efter hjärtstoppet. Nedkylning används i hopp om att förbättra överlevnadschanser samt det neurologiska utfallet för patienten. Därför analyseras dödlighet samt neurologisk utfall definierad som värdet på Cerebral Performance Categories (CPC) skalan.

För att få fram gruppindelningen har fyra olika metoder använts. Eftersom det är okänt hur många grupper som faktiskt finns är valideringen av resultaten mycket viktig. Utfallsvariabler såsom patienternas CPC värde är kända och kan användas för valideringen. Gruppindelningen tillsammans med TTM variabeln kan sedan användas för att förutspå patientens neurologiska utfall där utfallet kan vara bra, vilket innebär ett CPC värde mellan 1 och 3, eller dåligt om CPC värdet är 4 eller 5 där 5 betyder att patienten har avlidit.

De fyra olika metoderna leder till olika indelningar där både antalet grupper och indelningen av patienterna varierar. För det första dataset varierar antalet grupper mellan två och sex medan det bara varierar mellan två och tre grupper för det andra dataset. Att dra slutsatser från detta arbete är möjligt eftersom det finns en metod, KAMILA, vars indelning leder till de bästa resultaten för båda dataseten. Skillnader i både mortalitet och neurologisk utfall mellan olika grupper kan observeras när dataseten grupperas med KAMILA. Även prediktionen av patienternas neurologiska utfall leder till lovande resultat där det bland annat tyder på att nedkylningen har större effekt på patienter i vissa grupper.

Abstract

In this thesis, it is analysed if cardiac arrest patients can be grouped into similar clusters based on different underlying conditions and clinical variables and if there is a difference between clusters in either mortality or neurological outcome as measured by the Cerebral Performance Categories (CPC) scale. The two data sets both contain a targeted temperature management variable which indicates whether or not patients are cooled down upon arrival as well as a variety of continuous and categorical variables. Thus, the clustering methods need to be able to handle mixed data. The four methods that are presented in this thesis are Latent Class Analysis, KAMILA, which stands for KAy-means for MIxed LArge data, k-prototypes, and Partitioning Around Medoids with Gower's distance. These methods are then applied to the two data sets of cardiac arrest patients in order to find underlying phenotypes. For both data sets, when only using the cluster assignment and targeted temperature management variables to predict the binary CPC score, the KAMILA algorithm leads to the best results. Furthermore, there is also a significant difference in CPC score and mortality across the obtained clusters. The evidence suggests that it is not only possible to cluster cardiac arrest patients into different groups based on variables obtained upon admission and the patients' medical history but also that the cooling might be more useful to some clusters than others.

Contents

1 Introduction

According to the American Heart Association [1], cardiac arrest is defined as "the abrupt loss of heart function in a person who may or may not have been diagnosed with heart disease". There are many different causes for cardiac arrest such, as an abnormal heart rhythm, called ventricular fibrillation, as well as underlying heart diseases [2]. Often cardiac arrest leads to severe brain injury or even death.

In more recent years, different unsupervised learning methods have been applied to a multitude of sepsis patient populations with the goal of finding distinct subgroups. For example, Self-Organizing Maps (SOM) have been used to identify sepsis clusters that might lead to personalizing the treatment in the future [3] and Latent Class Analysis (LCA) has been used to find out whether sepsis, a quite heterogeneous sickness, stems from different underlying conditions which might lead to a difference in mortality [4]. Since the results have been promising when studying sepsis and not much research on clustering of cardiac arrest patients has been done, applying unsupervised learning methods to other heterogeneous conditions, such as cardiac arrest, is of great interest.

Data sets that contain both numeric and categorical variables, often called mixed data, are very common not only in the health domain, where the assessment of medical conditions and prediction of outcomes is important but also in biology, finance and marketing [5]. The main problem when dealing with mixed data is that the majority of unsupervised clustering algorithms is based on a distance measure to calculate the similarity of different observations. For example, when only having continuous variables, the Euclidean distance is commonly used. While it is possible to calculate distance measures both for continuous and categorical variables separately, it is not straightforward how to combine them since similarity is not well defined for mixed data. This leads to different algorithms that can deal with mixed data such as Partitioning Around Medoids (PAM), [6], as well as k-prototypes, which is an adaption of the widely used k-means algorithm, which can be used with mixed data [7]. Another algorithm related to k-means clustering is KAMILA, which can handle large data sets containing both continuous and categorical variables [8].

LCA is widely used in health research due to its ability to handle continuous and categorical variables at the same time [9]. It is a probabilistic model where each individual is assigned a probability of belonging to each subgroup or cluster. Here, the assumption is that each subgroup comes from a different underlying distribution.

Independent of which algorithm is used, individuals of two different subgroups should be as dissimilar as possible. Finding different clusters within a patient population can then lead to a better understanding of the individuals within each group and can possibly lead to distinct clinical profiles as well as a better understanding of the underlying phenomenon [10].

1.1 Thesis Objective

The aim of this thesis is to find subgroups within cardiac arrest patients that share similar characteristics and to evaluate whether those clusters have a significant effect on either mortality or neurological outcome as measured by the Cerebral Performance Categories (CPC) scale.

2 Theory

2.1 Finite Mixture Models

Given the observations $\boldsymbol{x}$, the general mixture model, as stated in [11], is defined as

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m \phi(\boldsymbol{x}; \boldsymbol{\theta}_m) \tag{1}$$

where $\phi(\boldsymbol{x}; \boldsymbol{\theta}_m)$ is the density depending on some parameters $\boldsymbol{\theta}_m$, possibly vector-valued. Furthermore, α_m are the mixing proportions where we assume that

$$\sum_{m=1}^{M} \alpha_m = 1.$$

Assuming that the density is Gaussian and thus setting $\phi(\boldsymbol{x}; \boldsymbol{\theta}_m) = \phi(\boldsymbol{x}; \mu_m, \Sigma_m)$ in Equation (1), we obtain the widely used Gaussian mixture model. Independent of which density is used, the parameter estimates $\hat{\boldsymbol{\theta}}_m$ can be obtained by maximum likelihood estimation.

Given the parameter estimates $\hat{\boldsymbol{\theta}}_m$, the mixture model can then also be used for classification. The probability that observation i belongs to class m can be estimated as

$$\hat{r}_{im} = \frac{\hat{\alpha}_m \phi(x_i; \hat{\boldsymbol{\theta}}_m)}{\sum_{k=1}^{M} \hat{\alpha}_k \phi(x_i; \hat{\boldsymbol{\theta}}_k)}$$

which is then used for classification.

2.2 Latent Class Analysis (LCA)

The traditional latent cluster model, as first introduced by Goodman in 1974, groups nominal variables into T different classes where individuals within each class should be as homogeneous as possible. On the other hand, individuals from different classes should be as heterogeneous as possible. The clusters are given by categorical, latent variables, and each individual or case is only assigned to one class [12]. A key assumption of Latent Class Analysis (LCA), which is a special case of finite mixture models [10], is that of local independence between variables belonging to different classes. Thus, all similarities in variables are assumed to be explained by the latent variable.

Given r nominal variables $u_1, \ldots, u_r$, the probability of belonging to class k is given by

$$P(u_1, \ldots, u_r, k) = P(c = k) \prod_{i=1}^{r} P(u_i | c = k) \tag{2}$$

where $P(c = k)$ denotes the probability of being in class k, for $k = 1, \ldots, K$, and $P(u_i | c = k)$ is the conditional probability of obtaining response u_i [12, 13].

From Equation (2), the joint probability can be obtained as

$$P(u_1, \ldots, u_r) = \sum_{k=1}^{K} P(c = k) \prod_{i=1}^{r} P(u_i | c = k). \tag{3}$$

Classification of individuals is based on the posterior probability. Using Bayes' theorem as well as replacing the model parameters by their maximum likelihood estimates in Equations (2) and (3), the posterior probability of being in class k becomes

$$\hat{P}(c = k | u_1, \ldots, u_r) = \frac{\hat{P}(u_1, \ldots, u_r, k)}{\hat{P}(u_1, \ldots, u_r)}.$$

Thus, each individual is assigned to the class for which it has the highest posterior probability.

While the traditional latent cluster model is based on nominal variables, this assumption can be lifted. Given r continuous responses of individual i, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ir})$, Equation (3) can be written as

$$f(\boldsymbol{y}_i) = \sum_{k=1}^{K} P(c = k) f(\boldsymbol{y}_i | c = k)$$

where, as previously, c denotes the categorical latent class variable with $k = 1, \ldots, K$ and $f(\boldsymbol{y}_i | c = k)$ is the multivariate normal density where both the mean and the variance possibly are class-specific [13]. The posterior probability of being in class k can be obtained in a similar way as for the nominal responses.

Furthermore, even a mixture of nominal and continuous variables can be modelled with LCA.

One of the key advantages of LCA is that the parameters can be estimated using maximum likelihood estimation. Thus, model fit measures such as the log-likelihood as well as the Akaike Information Criterion and Bayesian Information Criterion, see Section 2.3, can be calculated and used to evaluate how well the model fits the data [10]. This is important for model evaluation since the latent variable cannot be observed.

2.2.1 Classification error

In LCA, the classification error E is defined as

$$E = \frac{\sum_{i=1}^{I} w_i \Big(1 - \max \hat{P}(x | \boldsymbol{z}_i, \boldsymbol{y}_i)\Big)}{N}$$

where x is the cluster variable, $\boldsymbol{y}_i$ are the response variables, also called indicator variables, and $\boldsymbol{z}_i$ the exogenous variables, also called covariates. Furthermore, we have that

$$\hat{P}(x | \boldsymbol{z}_i, \boldsymbol{y}_i) = \frac{\hat{P}(x | \boldsymbol{z}_i) \hat{f}(\boldsymbol{y}_i | x, \boldsymbol{z}_i)}{\hat{f}(\boldsymbol{y}_i | \boldsymbol{z}_i)},$$

$$N = \sum_{i=1}^{I} w_i$$

where N is the number of observations, I the number of cases and w_i the case weight [14]. We want the classification error to be as close to zero as possible, since it estimates how many observations are misclassified, i.e., how many observations are assigned to the wrong class.

2.3 Information Criteria

The Akaike Information Criterion (AIC) is defined as

$$AIC = -2 \log(L) + 2p$$

where L denotes the maximum likelihood and p the number of estimated parameters [11]. Similarly, the Bayesian Information Criterion (BIC) is defined as

$$BIC = -2 \log(L) + p \log(n)$$

where n is the sample size. For both criteria, the idea is that we minimize a penalized likelihood function in order to avoid over-fitting the model, where a lower value thus indicates a better model fit. Here, it can also be mentioned that the BIC usually leads to simpler models, since it penalizes larger models more.

2.4 k-means

One of the most popular clustering algorithms for continuous data is called k-means. The idea is to cluster a data set with n observations around $k \leq n$ cluster means [11, 7]. The goal is to minimise the sum-of-squared errors

$$\begin{aligned} SS &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p} (x_{i,j} - q_{l,j})^2 \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l} ||x_i - q_l||_2^2 \end{aligned} \tag{4}$$

given the cluster means $q_{l,j}$ under the constraint that

$$\sum_{l=1}^{k} w_{i,l} = 1 \text{ for } 1 \leq i \leq n, \tag{5}$$

$$w_{i,l} \in \{0, 1\} \text{ for } 1 \leq i \leq n, \quad 1 \leq l \leq k. \tag{6}$$

This is described in more detail for the k-prototypes algorithm, which can be seen as a version of the k-means algorithm with mixed data, in Section 2.7. However, assuming that there are no categorical variables or setting $\lambda = 0$, you get the k-means algorithm when following Algorithm 2.

2.5 KAMILA

While k-means is an efficient algorithm to cluster large data sets, it can only be applied to continuous data. However, when dealing with mixed data, KAMILA, which stands for KAy-means for MIxed LArge data and is described in [15, 8], is becoming a popular alternative and can be seen as a combination of k-means and mixture models.

Assume that $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_i, \ldots \boldsymbol{V}_N)$ with $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{ip}, \ldots, V_{iP})^T$ is an i.i.d. sample of continuous vectors of dimension $P \times 1$. Furthermore, assume that $\boldsymbol{V}$ follows a finite mixture distribution of elliptical distributions such that

$$\boldsymbol{V}_i \sim f_{\boldsymbol{V},g}(\boldsymbol{v}) = \sum_{g=1}^{G} \pi_g h(\boldsymbol{v}; \boldsymbol{\mu}_g). \tag{7}$$

Here, $g = 1, \ldots, G$ denotes cluster membership, $\boldsymbol{\mu}_g$ the centroid of cluster g and π_g the prior probability of drawing an observation from cluster g. How to choose G is described in Section 2.6.

Similarly, let $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_i, \ldots \boldsymbol{W}_N)$ with $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iq}, \ldots, W_{iQ})^T$ be an i.i.d. sample of categorical vectors of dimension $Q \times 1$. Also assume that $\boldsymbol{W}$ follows a finite mixture distribution of multinomial distributions such that

$$\boldsymbol{W}_i \sim f_{\boldsymbol{W},g}(\boldsymbol{w}) = \sum_{g=1}^{G} \pi_g \prod_{q=1}^{Q} m(w_q; \boldsymbol{\theta}_{gq}) \tag{8}$$

where $m(w_q; \boldsymbol{\theta}_{gq})$ denotes the multinomial probability mass function with parameter vector $\boldsymbol{\theta}_{gq}$.

Under the assumption that $\boldsymbol{V}$ and $\boldsymbol{W}$ are locally independent, combining Equations (7) and (8) leads to the following joint density

$$\begin{aligned} &f_{\boldsymbol{V},\boldsymbol{W},g}(\boldsymbol{v}, \boldsymbol{w}) = f_{\boldsymbol{V},g}(\boldsymbol{v}) f_{\boldsymbol{W},g}(\boldsymbol{w}) \implies \\ &f_{\boldsymbol{V},\boldsymbol{W},g}(\boldsymbol{v}, \boldsymbol{w}) = \sum_{g=1}^{G} \pi_g f_{\boldsymbol{V},\boldsymbol{W},g}(\boldsymbol{v}, \boldsymbol{w}). \end{aligned} \tag{9}$$

The unknown parameters $\boldsymbol{\mu}_g$ and $\boldsymbol{\theta}_{gq}$ in Equation (9) are then estimated according to Algorithm 1.

Algorithm 1: KAMILA

1. Start by randomly initializing the parameter estimates. For all g and q, randomly draw from the observations of the continuous variables to obtain $\hat{\boldsymbol{\mu}}_q^{(0)}$ and draw from Dirichlet distribution with all concentration parameters $\alpha_i = 1$ to get $\hat{\boldsymbol{\theta}}_{gq}^{(0)}$.

2. Given $\hat{\boldsymbol{\mu}}_q^{(t)}$ and $\hat{\boldsymbol{\theta}}_{gq}^{(t)}$ repeat the following steps until convergence.

   (a) Calculate the Euclidean distance to each centroid $\hat{\boldsymbol{\mu}}_q^{(t)}$ according to

   $$d_{ig}^{(t)} = \sqrt{\sum_{p=1}^{P} \left(v_{ip} - \hat{\mu}_{qp}^{(t)}\right)^2}.$$

   (b) Find $r_i^{(t)} = \min_g d_{ig}^{(t)}$.

   (c) Given N observations, let $r = \sqrt{\boldsymbol{v}^T \boldsymbol{v}}$, $k(\cdot)$ be a Gaussian kernel and $h = 0.9An^{-1/5}$ be the bandwidth with $A = \min\left(\hat{\sigma}, \frac{\hat{q}}{1.34}\right)$, where $\hat{\sigma}$ denotes the sample standard variation and $\hat{q}$ the sample interquartile range. Then, calculate the kernel density estimate of the minimum distances

   $$\hat{f}_R^{(t)}(r) = \frac{1}{Nh^{(t)}} \sum_{l=1}^{N} k\Big(\frac{r - r_l^{(t)}}{h^{(t)}}\Big)$$

   which, according to Proposition 2 in [15], can be used to construct

   $$\hat{f}_{\boldsymbol{V}}(\boldsymbol{v}) = \frac{\hat{f}_R(r)\Gamma\left(\frac{P}{2} + 1\right)}{Pr^{P-1}\pi^{P/2}}$$

   where $\boldsymbol{V} = (V_1, \ldots, V_P)$.

   (d) Given cluster membership g, calculate the probability of observing the $\boldsymbol{W}_i$,

   $$c_{ig}(t) = \prod_{q=1}^{Q} m(w_{iq}; \hat{\boldsymbol{\theta}}_{gq}).$$

   (e) Assign observation i to cluster g such that

   $$H_i^{(t)}(g) = \log\left(\hat{f}_{\boldsymbol{V}}(d_{ig}^{(t)})\right) + \log\left(c_{ig}(t)\right)$$

   is maximized.

   (f) Let the set of indices of observations assigned to cluster g at iteration t be denoted by $\Omega_q^{(t)}$ and let $I(\cdot)$ be the indicator function. Re-estimate the parameters

   $$\hat{\boldsymbol{\mu}}_g^{(t+1)} = \frac{1}{|\Omega_q^{(t)}|} \sum_{i \in \Omega_g^{(t)}} \boldsymbol{v}_i,$$

   $$\hat{\theta}_{gql}^{(t+1)} = \frac{1}{|\Omega_q^{(t)}|} \sum_{i \in \Omega_g^{(t)}} I(w_{iq} = l).$$

3. Partition the observations such that $\sum_{i=1}^{N} \max_g \left(H_i^{\text{final}}(g)\right)$ is maximized.

2.6 Prediction Strength

As mentioned in [15], prediction strength is commonly used to decide upon cluster enumeration when using KAMILA and is described in [16].

In general, let $C(Y,k)$ be a clustering method that clusters a data set Y into k clusters. Furthermore, let Z be another data set. Then, $D\Big(C(Y,k),Z)_{ii'}\Big)$ is a symmetrical matrix where

$$D\Big(C(Y,k),Z)\Big)_{ii'} = \begin{cases} 1 & \text{if observation } Z_i \text{ and } Z_i' \text{ map to the same cluster obtained by } C(Y,k) \\ 0 & \text{otherwise} \end{cases}.$$

Now, assume that we have a training set X_{train} consisting of p features as well as a test set X_{test}. Furthermore, assume that they come from the same population. If a test set is not available, cross-validation is used instead.

Let A_{kj} be the indices of all observations of the test set being in cluster j and $n_{kj} = |A_{kj}|$. Then, the prediction strength $\text{ps}(k)$ is defined as

$$ps(k) = \min_{1\leq j\leq k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i\neq i'\in A_{kj}} D\Big(C(X_{\text{train}},k),X_{\text{test}}\Big)_{ii'}.$$

The prediction strength tells us how well the clustering of the training set predicts the clustering of the test set. Assume that there actually are k_0 clusters. If $k = k_0$, then $ps(k)$ should be close to 1 since the training set should be good at predicting the clusters of the test set. However, if $k > k_0$, the clusterings are much more dissimilar. Hence, we choose the number of clusters to be the largest k such that $ps(k) > T$ where T is some threshold. Commonly, $T = 0.8$ works well as long as the clusters are well separated [16].

2.7 k-prototypes

The k-prototypes algorithm can be seen as an extension of the k-means algorithm which can also be applied to mixed data and thus we cannot use Euclidean distance as the distance measure. As described in [17], the following distance measure is used for the k-prototypes algorithm

$$d(x_j,y_j) = \sum_{j=1}^{p}(x_j-y_j)^2 + \lambda \sum_{j=p+1}^{m} \delta(x_j,y_j). \tag{10}$$

Here, the first p variables are assumed to be continuous and thus the Euclidean distance can be used as a distance measure. The remaining $m-(p+1)$ variables are categorical and we thus use the following dissimilarity measure

$$\delta(x_j,y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j \end{cases}$$

which, similar to the Euclidean distance, has the property that the more similar x_j and y_j are, the smaller value it takes.

If the weighting parameter $\lambda = 0$, we obtain the k-means algorithm. The larger λ is, the more weight is put on the categorical variables in relation to the continuous ones. However, choosing λ is usually not straightforward since it varies depending on which variables are deemed more important. As stated in [17], a general estimate can be obtained as

$$\lambda = \frac{\sigma}{h_{cat}}$$

where σ is the average standard deviation of the continuous variables and

$$h_{cat} = \frac{1}{m-(p+1)} \sum_{j=p+1}^{m} \big(1 - \sum_i p_{ji}^2\big),$$

where p_{ji} denotes the probability that variable j takes on value i.

A more detailed explanation of the algorithm, as outlined in [7], follows below. Assume that we have n observations $\boldsymbol{X} = (X_1, \ldots, X_n)$, where

$$X_i = [x_{i,1}, \ldots, x_{i,p}, x_{i,p+1}, \ldots, x_{i,m}].$$

Thus, the data set can be described by the following m attribues, $A_1, \ldots, A_m$. Furthermore, assume that we have k clusters, $1 \leq k \leq n$, a $n \times k$ partition matrix W with the same constraints as given in Equations (5) and (6) and a cluster assignment matrix $\boldsymbol{Q} = \{Q_1, \ldots, Q_k\}$, where Q_l denotes the clustering center of cluster l for $1 \leq l \leq k$.

The goal of the k-prototypes algorithm is then to minimise the following cost function

$$P(W, \boldsymbol{Q}) = \sum_{l=1}^{k} \Big(\underbrace{\sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p} (x_{i,j} - q_{l,j})^2}_{P_l^n} + \lambda \underbrace{\sum_{i=1}^{n} w_{i,l} \sum_{j=p+1}^{m} \delta(x_{i,j}, q_{l,j})}_{P_l^c} \Big) \tag{11}$$

$$= \sum_{l=1}^{k} (P_l^n + P_l^c) \tag{12}$$

for $1 \leq l \leq k$.

The minimization is achieved by iterating steps 1 and 2 of Algorithm 2 until convergence and thus finding a local optimum.

Algorithm 2: k-prototypes

1. Given some $\hat{\boldsymbol{Q}}$, use Equation (10) to compute W such that

$$\begin{cases} w_{i,l} = 1, \text{ if } d(X_i, Q_l) \leq d(X_i, Q_t) \\ w_{i,t} = 0, \text{ if } t \neq l \end{cases}.$$

2. Given $\hat{W}$, compute $\boldsymbol{Q}$ by minimizing Equation (11). Since both P_l^n and P_l^c are non-negative, this is equivalent to minimizing Equation (12).

   (a) The cost function for the numeric variables, P_l^n for $1 \leq l \leq k$, is minimised by

   $$q_{l,j} = \frac{\sum_{i=1}^{n} w_{i,l} x_{i,j}}{\sum_{i=1}^{n} w_{i,l}} \text{ for } 1 \leq j \leq m.$$

   (b) The cost function for the categorical variables P_l^c for $1 \leq l \leq k$, is minimised by choosing q_j such that the relative frequency of q_j is greater than the relative frequency of all other categories $c_{k,j}$. This can be written as

   $$f_r(A_j = q_j | \boldsymbol{X}) \geq f_r(A_j = c_{k,j} | \boldsymbol{X})$$

   for $p + 1 \leq j \leq m$ and $c_{k,j} \neq q_j$.

3. Repeat steps 1 and 2 until convergence.

2.8 Partitioning Around Medoids (PAM)

An algorithm similar to k-means is called k-medoids, and a short outline can be found in [6]. Using Euclidean distance as a distance measure, Equation (4) can be rewritten as

$$TD = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} ||x_i - \mu_l||_2^2. \tag{13}$$

where TD stands for "total deviation" and μ_l is an observation of cluster l. Thus the difference to k-means is that μ_l is an existing data point whereas $q_l \in \mathbb{R}^m$.

More general, when using any distance measure $d(\cdot, \cdot)$, Equation (13) can be written as

$$TD = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} d(x_i, \mu_l). \tag{14}$$

When using mixed data a common distance measure is Gower's distance [18], which is defined as

$$d(i,j) = \frac{1}{p} \sum_{k=1}^{p} d_{ij}^{(k)}.$$

If feature k is continuous, then

$$d_{ij}^{(k)} = 1 - \frac{|x_i - x_j|}{R_k},$$

where R_k denotes the range of feature k. On the other hand, if feature k is categorical, then

$$d_{i,j}^{(k)} = \begin{cases} 1, & x_i = x_j \\ 0, & x_i \neq x_j \end{cases}.$$

The Partitioning Around Medoids, or short PAM, algorithm is one of the most common algorithms to find clusters around k-medoids, [6]. The first part of the algorithm is called BUILD and gives us the initial clustering. In the second part of the algorithm, SWAP, the medoids are changed until a local optimum is found.

Algorithm 3: PAM algorithm

1. BUILD
   (a) Choose the first medoid as the observation with the least distance to all other observations.
   (b) Add the remaining $k-1$ medoids one by one. Choose the observation that minimises Equation (14) as medoid.
2. SWAP
   (a) For each medoid, check if there are any non-medoids that would lead to a smaller TD. If this is the case, the medoid is chosen as the observation that reduces TD the most.
   (b) Repeat until convergence.

2.9 Silhouette Width

The average silhouette width is commonly used when deciding on the number of clusters when for example using k-prototypes or PAM. In for example [19], it is defined in the following way: Assume that we have

k clusters and that $d(i, j)$ is the distance between two observations i and j. Given that observation i is in cluster C_i, then

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j).$$

Thus, a_i denotes the the average distance to other observations in the same cluster. On the other hand, b_i denotes the smallest average distance to any of the remaining $k - 1$ clusters. So

$$b_i = \min_l \Big(\frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \Big)$$

where $l = 1, \ldots, i - 1, i + 1, \ldots k$. Then, the average silhouette width is the average over all s_i with

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}.$$

If s_i is close to zero, this implies that it is unclear whether observation i should be assigned to cluster C_i or its neighbouring cluster. Thus, we want s_i to be as close to one as possible, since this implies that observation i has been assigned to right cluster, namely cluster C_i.

2.10 Cluster Validity

When performing unsupervised learning, the correct number of clusters is not known, and thus one needs to try to validate the obtained results.

2.10.1 Fisher's Exact Test

Fisher's exact test is a commonly used alternative to the Chi-squared test, especially when either the sample size is small or if there are too few observations in some of the groups [20]. It is also used to determine whether two groups are independent. Fisher's exact test is based on the hypergeometric distribution.

Table 1: Contingency table.

	A	B	Total
Group 1	a	b	r_1
Group 2	c	d	r_2
Total	c_1	c_2	N

Assume that you have 2×2 contingency table such as Table 1. Here, a, b, c, and d are integers and, for example, a denotes the number of observations that are in Group 1 and meet criterion A. The probability of obtaining that table is

$$p_{\text{table}} = \frac{\binom{c_1}{a}\binom{c_2}{b}}{\binom{N}{r_1}} = \frac{c_1! c_2! r_1! r_2!}{N! a! b! c! d!}, \tag{15}$$

as described in [21]. The probability given in Fisher's exact test is then obtained by repeating the calculations of Equation (15) for all combinations of a, b, c and d leading to c_1, c_2, r_1 and r_2 as in the original table, Table 1, and summing up all p-values less than or equal to p_{table}. This can also be extended to tables bigger than 2×2.

2.10.2 Rand Index

The Rand index is commonly used to compare different clusters. In supervised learning, it can be used to compare the obtained clusters to the true clusters whereas in unsupervised learning it can be used to compare two clusters obtained from different methods. The Rand index is defined as

$$RI = \frac{a + d}{L} = \frac{a + d}{N(N - 1)/2}$$

where $L = a+b+c+d$ and N is the total number observations, [22]. Furthermore, denote the two different cluster assignments, where one of them can be the true cluster assignment, as $C^{(1)} = \{C_1^{(1)}, \ldots, C_n^{(1)}\}$ and $C^{(2)} = \{C_1^{(2)}, \ldots, C_m^{(2)}\}$. Then, we have the following

$$\begin{cases} a: & \text{all observations that belong to the same cluster of both } C^{(1)} \text{ and } C^{(2)} \\ b: & \text{all observations that belong to the same cluster of } C^{(1)} \text{ but different clusters of } C^{(2)} \\ c: & \text{all observations that belong to different cluster of } C^{(1)} \text{ but the same clusters of } C^{(2)} \\ d: & \text{all observations that belong to different clusters of both } C^{(1)} \text{ and } C^{(2)} \end{cases}.$$

2.10.3 Logistic Regression

Logistic regression, in its most simple form, is widely used when dealing with a binary response variable, i.e. $Y \in \{0, 1\}$. Assume that

$$p_i = P(Y_i = 1) = 1 - P(Y_i = 0)$$

and that $\boldsymbol{X} = (X_1, \ldots, X_p)$ are the independent variables. Then the simple logistic regression model, as for example given in [11], is defined as

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \boldsymbol{x}_i\boldsymbol{\beta} \tag{16}$$

for $i = 1, \ldots, n$. We obtain the probabilities p_i by rewriting Equation (16),

$$p_i = \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}}}.$$

Since $Y_i \sim Bin(1, p_i)$, the parameter estimates $\hat{\boldsymbol{\beta}}$ can be found by maximizing the likelihood function,

$$\begin{aligned} L(\boldsymbol{\beta}; \boldsymbol{Y}) &= P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i) \\ &= \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = \prod_{i=1}^{n} \frac{e^{Y_i\boldsymbol{x}_i\boldsymbol{\beta}}}{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}, \end{aligned}$$

which is equivalent to maximizing the log-likelihood function

$$l(\boldsymbol{\beta}; \boldsymbol{Y}) = \ln\Big(L(\boldsymbol{\beta}; \boldsymbol{Y})\Big) = \sum_{i=1}^{n} \Big(Y_i\boldsymbol{x}_i\boldsymbol{\beta} - \ln\big(1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}}\big)\Big).$$

However, since this has no closed-form solution, the Newton-Raphson algorithm is used to find the $\boldsymbol{\beta}$-estimates.

The odds ratio, OR, is a commonly reported value when performing logistic regression. The odds ratio of variable j, for $j = 0, \ldots, p$, is defined as $\text{OR}_j = e^{\beta_j}$. It tells us how the outcome is affected if we increase x_j by one unit [23]. Thus, given $\Delta x_j = 1$, we have that if

$$\begin{cases} \text{OR}_j < 1, & \text{the odds decrease} \\ \text{OR}_j = 1, & \text{the odds are not affected} \\ \text{OR}_j > 1, & \text{the odds increase} \end{cases}.$$

3 Method

3.1 Data

The four different clustering methods are applied to two different data sets containing different clinical variables. From the first data set, from this time forward referred to as the TTM data set, 935 observations and 67 variables are used for clustering. Here, modifications are made since some of the continuous variables have categorical entries. For example, the variable measuring the number of defibrillations has values within $[0, 20]$ as well as some observations which have the value "> 20". All those are then set to 20. Similar modifications are made for the lactate, pO_2, and pCO_2 variables. Furthermore, all observations with more than 20% of the clinical variables missing are removed. Here it should be noted that insulin is the variable with the most missing values, which is due to the fact that every person without diabetes has no value for insulin. Before imputing all missing values, this is altered such that if a person does not have diabetes, then the insulin variable is set to zero. This is done to avoid problems when imputing, such as more people using insulin than there are people that have diabetes.

From the second data set, the INTCAR data set, 4261 observations but only 53 variables are used for clustering. As previously, all observations where more than 20% of the variables are missing are removed. An exact list of which variables are included in each data set can be found in Appendix A and B.

3.2 Missing Data

The more classical clustering methods such as KAMILA, k-prototypes and PAM with Gower's distance do not work with missing data, so that there is a need for handling missing values. One option is to remove all observations with missing values. However, if there are many missing values, then this can drastically reduce the size of the data set and thus impact the results. For example, after omitting the outcome variables and removing the observations where more than 20% of the variables are missing, out of 935 observations, only 46 are complete in the TTM data set. On the other hand, there is no guarantee that data imputation replaces the missing value with the correct value, which also can affect the results. The imputation method that is used is k-nearest neighbours with $k = 10$. So in order to assess how big of an effect imputation has on the results, a simulation with the TTM data set is performed.

After adjusting the insulin variable, there are 572 observations with missing values left. To check the robustness of the clustering methods, the idea is to add noise to the imputed values to see how sensitive the different clustering methods are to small variations. This is straightforward for continuous variables where we can just draw from a uniform distribution and add between -10% and 10% of noise to the imputation. However, adding $\pm 10\%$ noise to the categorical variables has no meaning. Thus, we replace the missing values by drawing from the distribution of each variable. While this is not equivalent to adding noise, it ensures some variation in the imputation. This is repeated ten times and each time compared to the original clustering obtained when imputing with k-nearest neighbours, and both the Rand Index and the Adjusted Rand Index is calculated. The results can be found in Table 2.

Table 2: Checking the robustness of clustering methods by adding some randomness to the imputed values and comparing it to the original clustering. The values reported for both the Rand Index and the Adjusted Rand Index are averages over ten repetitions with the standard deviation given in parentheses.

Clustering Method	**Rand Index**	**Min. Rand Index**	**Max. Rand Index**	**Adjusted Rand Index**
KAMILA	0.9734 (0.0058)	0.9622	0.9830	0.9467 (0.0117)
PAM	0.9506 (0.0074)	0.9459	0.9705	0.8996 (0.0150)
k-prototypes	0.9225 (0.0742)	0.7728	0.9725	0.8430 (0.1497)

As already implied by the high values of the Rand Index, for the majority of the 935 observations, the eleven cluster assignments, i.e., the original one as well as ten noisy ones, are identical. Interestingly

though, not only the observations with missing values, which thus are affected by adding noise to the imputation, change cluster assignment. This is illustrated in Table 3.

Table 3: The number of observations that are not assigned to the same cluster over all eleven repetitions as well as the number of complete observations that change cluster assignment.

Clustering Method	**Different cluster**	**Complete observations**
KAMILA	40 (4.28%)	13 (32.5%)
PAM	31 (3.32%)	3 (9.68%)
k-prototypes	161 (17.22%)	44 (27.33%)

3.3 Robustness of Clustering Methods

To ensure that the clustering results are reliable, it is common to use multiple initialisations. For both KAMILA and k-prototypes, 100 initialisations are used and we want to see if repeated clustering on the same data leads to identical clusters. This is repeated ten times, and the results for this can be found in Table 4.

Table 4: Checking the robustness of clustering methods by applying the same clustering algorithm on the same data and comparing clustering results to the initial, i.e. original, clustering. The values reported for both the Rand Index and the Adjusted Rand Index are averages over ten repetitions with the standard deviation given in parentheses.

Clustering Method	**Rand Index**	**Min. Rand Index**	**Max. Rand Index**	**Adjusted Rand Index**
KAMILA	0.9953 (0.0033)	0.9915	1	0.9906 (0.0066)
k-prototypes	0.9634 (0.0797)	0.77443	1	0.9261 (0.1609)

3.4 Latent Class Analysis

Since it is assumed that the continuous variables follow a Gaussian distribution, a transformation is applied where needed. This can easily be done since none of the positively skewed variables contains negative values. The transformation is either the logarithm if the variable does not contain zeros and is highly skewed. Otherwise, the square root is applied.

When fitting a latent cluster model, we start off by fitting a model with only one class, H_0 [12]. We then continuously add one class until a good model fit is obtained. However, it is not always straightforward whether a good model fit has been obtained. While finite mixture models such as LCA are widely used data analysis techniques, there is still no consistent measure used for deciding upon class enumeration. However, a different number of classes can lead to different groupings of the population and thus can greatly affect the results. Based on a Monte Carlo simulation study, [13], the BIC seems to be the best information criterion when trying to identify the correct number of classes and performs better than the AIC, the adjusted BIC and the consistent AIC. Thus, the number of clusters is based on choosing the model with the lowest BIC. However, there is nothing that ensures the convergence of the BIC or any other information criterion. So if the BIC does not converge, i.e. if the BIC continuously decreases as the number of clusters increases, we do not choose the model with the smallest BIC but look at the plot of the BIC and try to find where the BIC starts to decrease slower. This is commonly referred to as the "elbow criterion" [10].

An additional problem during model selection is that there is also no guarantee that the likelihood function converges to a global maximum [10]. In order to avoid convergence to a local maximum, 1000 random start values are used to avoid getting stuck in a local minimum.

3.5 KAMILA

Exactly as for LCA, either the logarithm or the square root is applied to the continuous variables that are highly skewed. While KAMILA does not assume that the continuous variables follow a Gaussian distribution, it is assumed that they follow an elliptical distribution [8]. While this is a less strict assumption, it is still assumed that elliptical distributions are not skewed [24].

In accordance with [8], all continuous variables are Z-transformed before clustering. Also, 100 initialisations are used to avoid getting a local solution. The number of clusters k is then chosen as the largest k where the prediction strength is still larger than some threshold [16]. Assuming that the clusters are well separated, a threshold value of 0.8 has been shown to work well to find the optimal k.

3.6 k-prototypes

Before clustering the data according to k-prototypes, all continuous variables are Z-scaled so that the clustering is not affected by on which scales those variables are measured. Once again, 100 initialisations are used to avoid local solutions. Then the number of clusters k is chosen such that the average silhouette width is maximized. Furthermore, it is tested whether different values of the weighting parameter λ lead to different k.

3.7 PAM

As for k-prototypes, the number of clusters is chosen to be the one that maximizes the average silhouette width. However, the scaling of the continuous variables is different. They are scaled to be within $[0, 1]$ instead of being Z-scaled.

4 Results

In this section, the results for each data set obtained by the four different clustering algorithms LCA, KAMILA, k-prototypes and PAM with Gower's distance are presented. These results include the appropriate measure used for cluster enumeration and how mortality, as well as the CPC score, vary over different clusters.

4.1 TTM data set

The following sections contain the results of the TTM data set. For more information, such as how the different clinical variables are distributed among the clusters, see Appendix A.

4.1.1 LCA

When applying LCA to the TTM data set, the BIC reaches its minimum for six clusters. See Table 5. Furthermore, as shown in Table 6, this leads to no cluster containing less than 5% of patients.

Table 5: The log-likelihood statistics as well as the classification error obtained when applying LCA to the TTM data set. The information criterion minima are marked in blue.

Clusters	Log-likelihood	BIC	AIC	Classification Error
1	-48999	98641	98186	0.0000
2	-45054	91401	90486	0.0007
3	-42521	86984	85609	0.0006
4	-41529	85650	83815	0.0012
5	41002	85246	82952	0.0165
6	-40591	85075	82320	0.0198
7	-40290	85122	81908	0.0200

Table 6: The number of patients per cluster for the optimal solution when applying LCA to the TTM data set. The percentage is given in parentheses.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
270 (29)	199 (21)	190 (20)	156 (17)	60 (6)	60 (6)

In Figure 1, it can be seen that the CPC score varies across the different clusters. For example, the majority of patients in the third cluster have a CPC score of 5, which is not only the worst outcome, it also implies that they are dead. On the other hand, almost 70% of the patients in the fourth cluster have a CPC score of either 1 or 2, which implies a good neurological outcome. The differences across different clusters are also confirmed by the results in Table 7. There is a significant difference in both mortality and CPC score.

Table 7: CPC score and mortality by cluster as determined by LCA. Fisher's exact test with simulated p-value is used to determine whether or not there is a significant difference. For the CPC score the five individuals with without score are removed for this analysis and the median as well as the Interquartile Range (IQR) are reported. If the p-value is below 0.05, it is marked by an asterisk, *.

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	p-value
180-day CPC score, median [IQR]	3 [1, 5]	1 [1, 5]	5 [5, 5]	1 [1, 5]	1 [1, 5]	5 [1, 5]	$< 0.001^*$
Mortality, n (%)	121 (45)	68 (34)	162 (85)	41 (26)	21 (35)	32 (53)	$< 0.001^*$

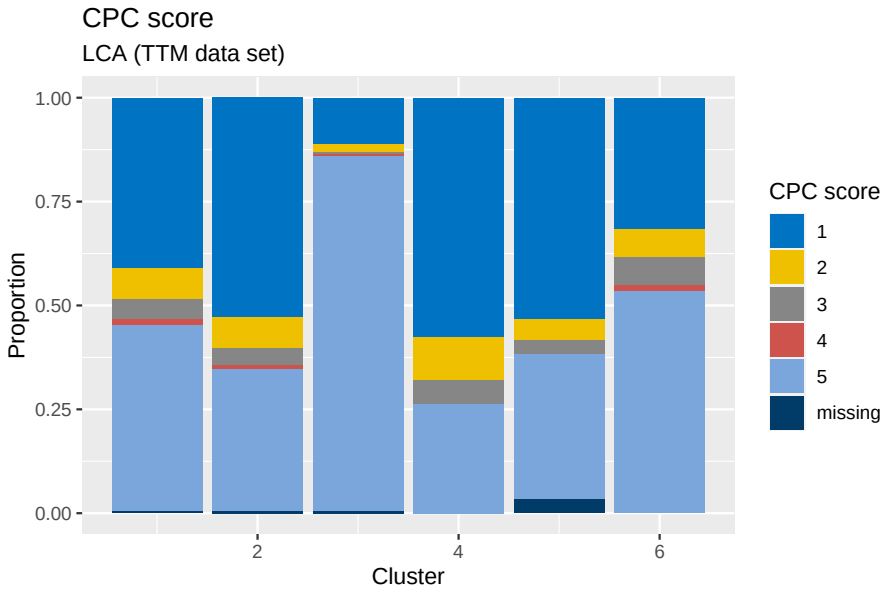

Figure 1: Distribution of the CPC score over different clusters obtained by LCA.

4.1.2 KAMILA

Figure 2 shows how prediction strength, the performance measure used for cluster enumeration when using KAMILA, varies as the number of clusters increases. A threshold of 0.8 or higher is commonly used for a well-separated cluster, and thus the data is clustered into two clusters which leads to the distribution stated in Table 8.

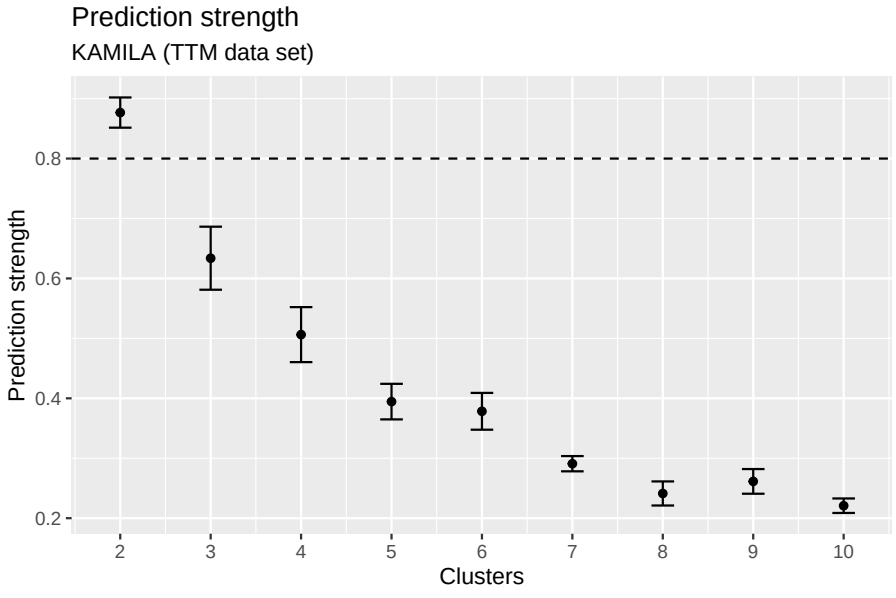

Figure 2: Prediction strength of the TTM data set for up to ten clusters where the error bars mark the standard errors. The dotted line marks the threshold of 0.8.

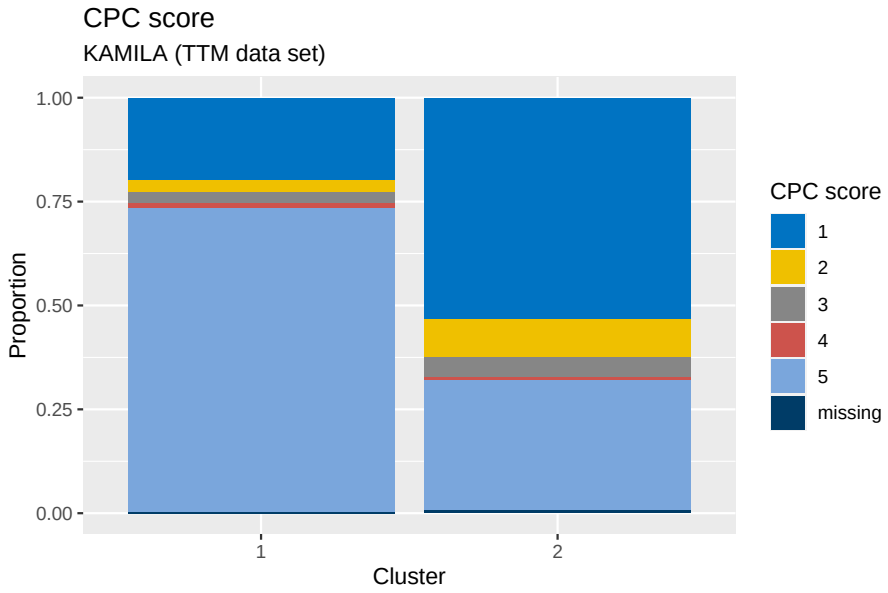

Figure 3: CPC scores of the two different clusters of the TTM data set obtained by KAMILA.

Table 8: The number of patients per cluster for the optimal solution when applying KAMILA to the TTM data set. The percentage is given in parentheses.

Cluster 1	**Cluster 2**
362 (39)	573 (61)

When clustering the data into two clusters, this results in the first cluster where around 75% of patients have a bad neurological outcome, i.e. a CPC score between 3 and 5, and the second cluster where the majority has a good neurological outcome. This is shown in Figure 3. Furthermore, the results in Table 9 confirm that there is a significant difference between the two clusters when looking at both mortality as well as the CPC score.

Table 9: CPC score and mortalitity of the KAMILA-clustering. Fisher's exact test, with simulated p-value where needed, is used to determine significance. For the CPC score the five individuals with missing scores are removed.

Variable	**Cluster 1**	**Cluster 2**	***p*-value**
180-day CPC score, median [IQR]	5 [3, 5]	1 [1, 5]	$< 0.001^*$
Mortality, n (%)	265 (73)	280 (31)	$< 0.001^*$

4.1.3 k-prototypes

The data is clustered into two clusters since this is where the average silhouette width reaches its maximum, see Figure 4. This leads to the number of patients being distributed as stated in Table 10.

Table 10: The number of patients per cluster for the optimal solution when applying k-prototypes to the TTM data set. The percentage is given in parentheses.

Cluster 1	**Cluster 2**
625 (67)	310 (33)

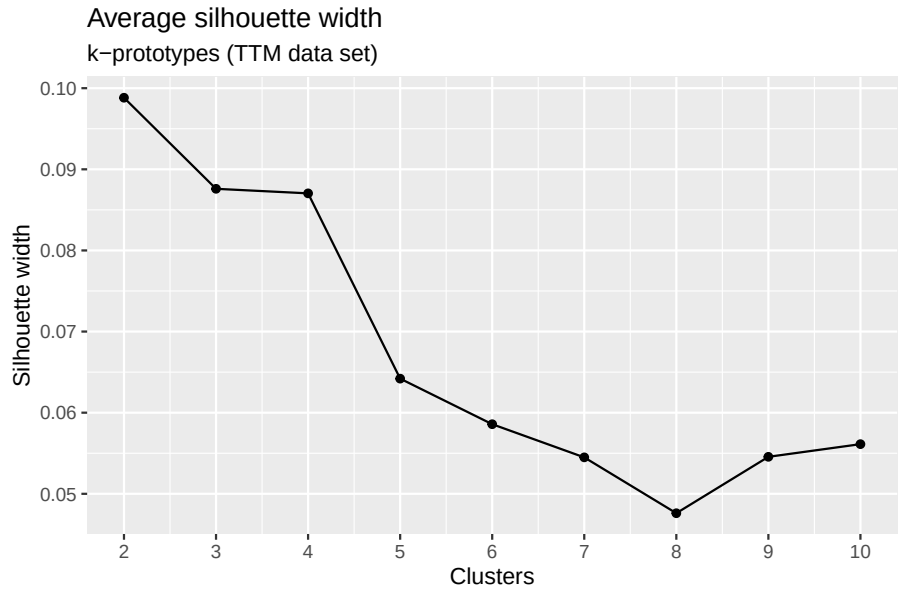

Figure 4: Average silhouette width when applying k-prototypes to the TTM data set.

While the distribution of CPC scores across the two clusters is quite similar, as shown in Figure 5, on average, the CPC score and the mortality are lower in the second cluster. As stated in Table 11, this difference is statistically significant.

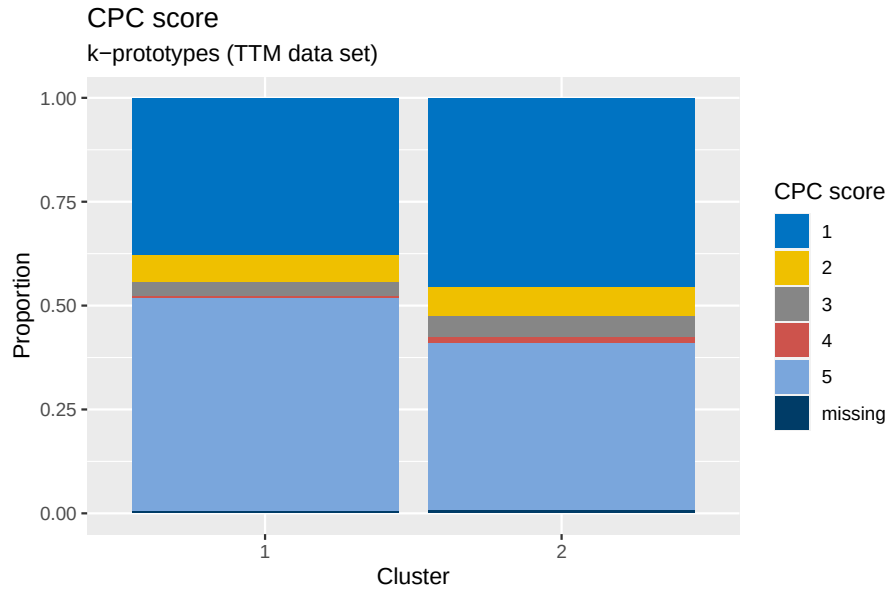

Figure 5: CPC scores over different clusters as obtained by k-prototypes.

Table 11: Outcome by cluster as determined by k-prototypes. Fisher's exact test, with simulated p-value where needed, is used to determine significance. For the CPC score the five individuals without any score are removed.

Variable	**Cluster 1**	**Cluster 2**	***p*-value**
180-day CPC score, median [IQR]	5 [1, 5]	2 [1, 5]	0.0255*
Mortality, n (%)	320 (51)	125 (40)	0.0018*

4.1.4 PAM

In accordance with theory, the data is clustered into two clusters since the average silhouette width reaches its maximum there, as shown in Figure 6. This leads to the patient distribution stated in Table 12.

Table 12: The number of patients per cluster for the optimal solution when applying PAM with Gower's distance to the TTM data set. The percentage is given in parentheses.

Cluster 1	**Cluster 2**
304 (33)	631 (67)

The distribution of the CPC score across the two clusters looks almost identical for both clusters, as can be seen in Figure 7. Furthermore, there is no significant difference in CPC score or mortality for the two clusters. See Table 13.

Table 13: CPC score and mortality by cluster as determined by PAM with Gower's distance. Fisher's exact test, with simulated p-value where needed, is used to determine significance. For the CPC score the five individuals without any score are removed.

Variable	**Cluster 1**	**Cluster 2**	***p*-value**
180-day CPC score, median [IQR]	3 [1, 5]	3 [1, 5]	0.9185
Mortality, n (%)	141 (46)	304 (48)	0.625

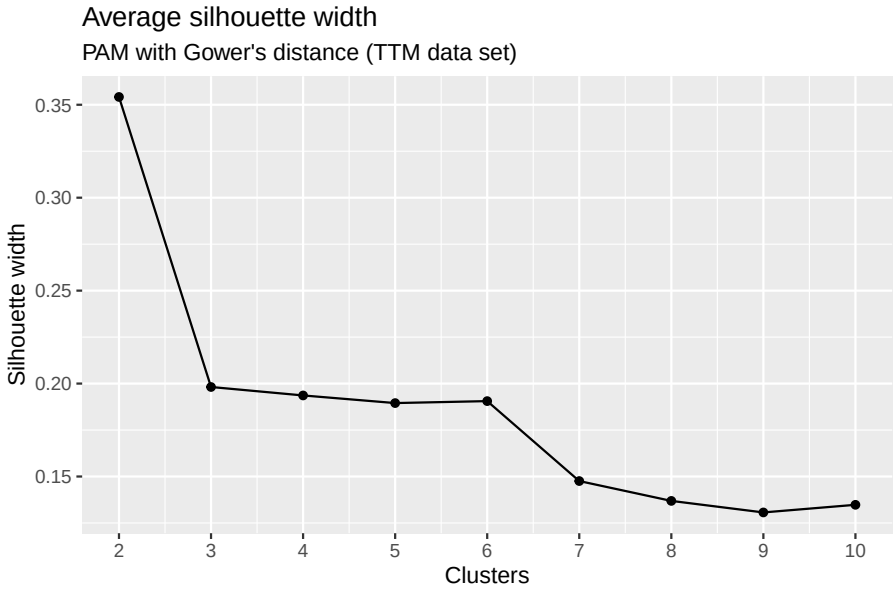

Figure 6: Average silhouette width for up to ten clusters when applying the PAM algorithm to the TTM data set.

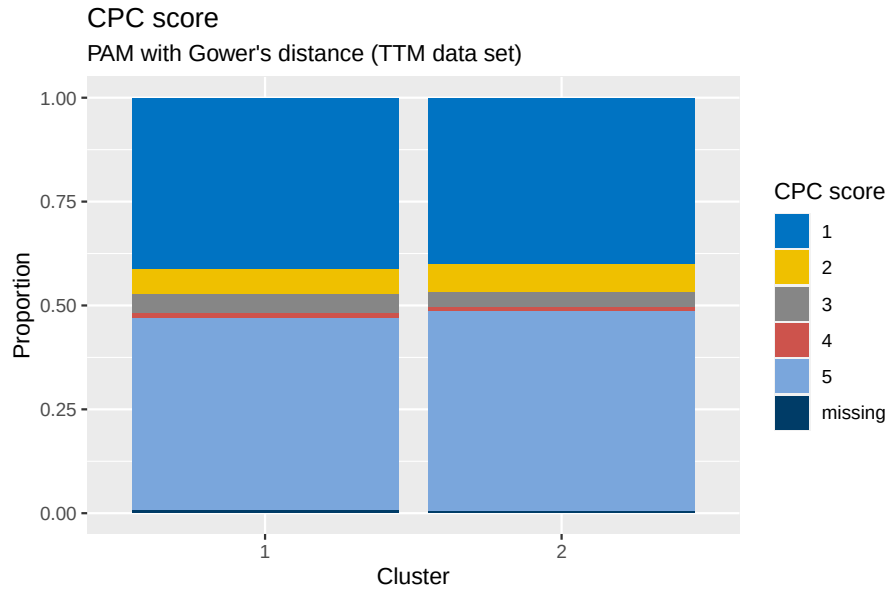

Figure 7: CPC score over different clusters obtained by PAM with Gower's distance for the TTM data set.

4.1.5 Logistic regression

We now want to try to predict the binary CPC score where a binary CPC score 0 means a good neurological outcome and 1 implies a bad neurological outcome. The binary CPC score is defined as

$$Y_{\text{binary CPC}} = \begin{cases} 0 & \text{if } Y_{\text{CPC}} \in \{1,2,3\} \\ 1 & \text{if } Y_{\text{CPC}} \in \{4,5\} \end{cases}. \tag{17}$$

The prediction is based on three different logistic regression models. The first one is the reference model containing variables that are medically significant variables that should affect the CPC score. The second model contains the cluster assignment variable as well as a variable indicating whether or not targeted temperature management, TTM, has been applied. For the clusters obtained by LCA, there is even another variation where the probabilities of belonging to each cluster instead of the cluster assignment variable are used. Lastly, the third model is a combination of the reference and the clustering model. We can also compute the area under the curve, AUC, as well as its corresponding confidence interval. Those results are given in Table 14 and 15.

Table 14: Comparing the area under the curve (AUC) of the reference model to the clustering as well as to the combined model where we try to predict the binary CPC score 180 days later. Here we test whether the AUC of the reference model is smaller than the AUC of the other model.

	LCA (Cluster membership)			**LCA (Cluster probabilities)**		
Model	**AUC**	**95% CI**	***p*-value**	**AUC**	**95% CI**	***p*-value**
Reference	0.8233	$(0.797 - 0.850)$	–	0.8233	$(0.797 - 0.850)$	–
Clustering	0.6704	$(0.636 - 0.704)$	1	0.6826	$(0.649 - 0.717)$	1
Combined	0.8345	$(0.809 - 0.860)$	0.01^*	0.8350	$(0.810 - 0.860)$	0.001^*

Table 15: Comparing the area under the curve (AUC) of the reference model to the clustering as well as to the combined model where we try to predict the binary CPC score. Here we test whether the AUC of the reference model is smaller than the AUC of the other model.

	KAMILA			*k*-prototypes			PAM		
Model	**AUC**	**95% CI**	***p*-value**	**AUC**	**95% CI**	***p*-value**	**AUC**	**95% CI**	***p*-value**
Reference	0.8233	(0.797 – 0.850)	–	0.8225	(0.796 – 0.849)	–	0.8225	(0.796 – 0.849)	–
Clustering	0.6965	(0.664 – 0.729)	1	0.5496	(0.514 – 0.585)	1	0.5176	(0.482 – 0.553)	1
Combined	0.8333	(0.808 – 0.859)	0.0148*	0.8224	(0.796 – 0.849)	0.5257	0.8247	(0.799 – 0.851)	0.212

The highest AUC-value for the clustering model is obtained by the KAMILA algorithm. The highest AUC-value for the combined model is obtained when using the cluster probabilities as given by LCA, followed by the model using the cluster assignment as given by LCA. When using the cluster assignment as given by KAMILA, we get the third highest AUC-value for the combined model. However, the difference between those three is only 0.0017.

Figure 8 shows the receiver operating characteristic curves, or short ROC curves, based on the KAMILA clustering. A list over which variables are included as well as their odds ratio, 95% confidence interval and *p*-value can be found in Table 16. Here, it is important to note that cluster assignment is a statistically significant variable.

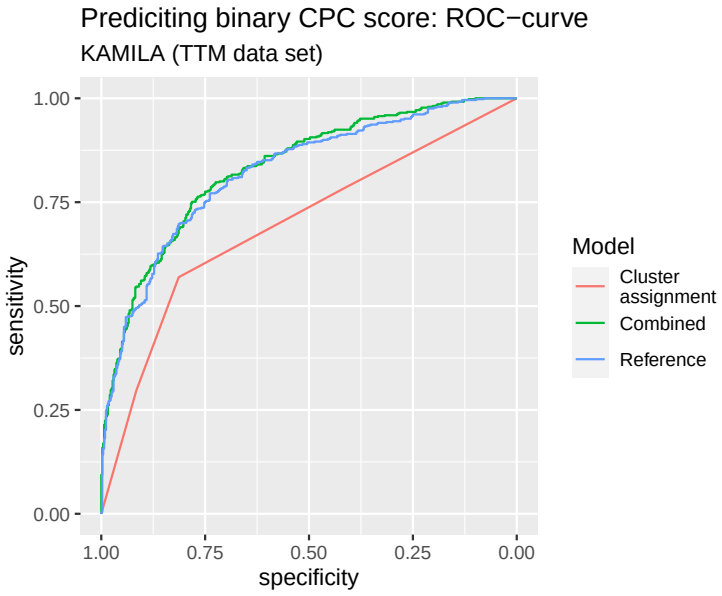

Figure 8: ROC-curves of the three different logistic regression models where the cluster assignment is obtained by applying KAMILA to the TTM data set.

Table 16: Odds ratio (OR) and the corresponding confidence interval (CI) of the reference model as well as of two clustering models obtained by KAMILA using the cluster assignment as factor variable. Here, TTM stands for target temperature management.

	Reference Model			Clustering Model KAMILA		
Variable	**OR**	**95% CI**	***p*-value**	**OR**	**95% CI**	***p*-value**
Intercept	13.38	$(6.63 - 27.91)$	$< 0.001^*$	3.92	$(2.76 - 5.70)$	$< 0.001^*$
Age	2.18	$(1.83 - 2.61)$	$< 0.001^*$	–	–	–
Sex	1.59	$(1.06 - 2.41)$	0.027^*	–	–	–
Bystander witnessed arrest	0.65	$(0.39 - 1.07)$	0.094	–	–	–
First rhythm shockable	0.18	$(0.11 - 0.30)$	$< 0.001^*$	–	–	–
ROSC	1.89	$(1.55 - 2.32)$	$< 0.001^*$	–	–	–
Adrenaline	1.19	$(1.00 - 1.42)$	0.056	–	–	–
GCS Motor	0.73	$(0.62 - 0.85)$	$< 0.001^*$	–	–	–
Pupillary reflex	0.39	$(0.26 - 0.58)$	$< 0.001^*$	–	–	–
Cluster 2	–	–	–	0.15	$(0.09 - 0.23)$	$< 0.001^*$
TTM	–	–	–	0.76	$(0.46 - 1.24)$	0.276
Cluster 2 · TTM	–	–	–	1.36	$(0.75 - 2.48)$	0.317

4.2 INTCAR data set

The following sections contain the results of the INTCAR data set. For more information, such as how the different clinical variables are distributed among the clusters, see Appendix B.

4.2.1 LCA

For the INTCAR data set, the BIC does not reach a minimum. See Table 17. Thus, the elbow criterion is used to conclude that the optimal number of clusters is three, as shown in Figure 9.

Table 17: The log-likelihood statistics as well as the classification error obtained when applying LCA to the INTCAR data set.

Clusters	**Log-likelihood**	**BIC**	**AIC**	**Classification Error**
1	-144965	290498	290066	0.0000
2	-126942	255029	254158	0.0095
3	-121906	245535	244225	0.0142
4	-120084	242466	240717	0.0131
5	-118540	239954	237767	0.0176
6	-117531	238514	235888	0.0262
7	-116156	236341	233277	0.0251
8	-115255	235114	231611	0.0261

This leads to the number of patients per cluster being distributed according to Table 18.

Table 18: Number of patients per cluster as obtained by LCA with the percentage given in parentheses.

Cluster 1	**Cluster 2**	**Cluster 3**
1995 (47)	1479 (35)	787 (18)

When looking at Figure 10 it can be seen that the majority of patients in the first cluster have a high CPC score. The distribution of CPC scores are quite similar in the second and third cluster, where the majority of the patients have a low CPC score and thus a good neurological outcome. According to the results in Table 19, the difference in distribution of both CPC score and mortality across the three clusters is significant.

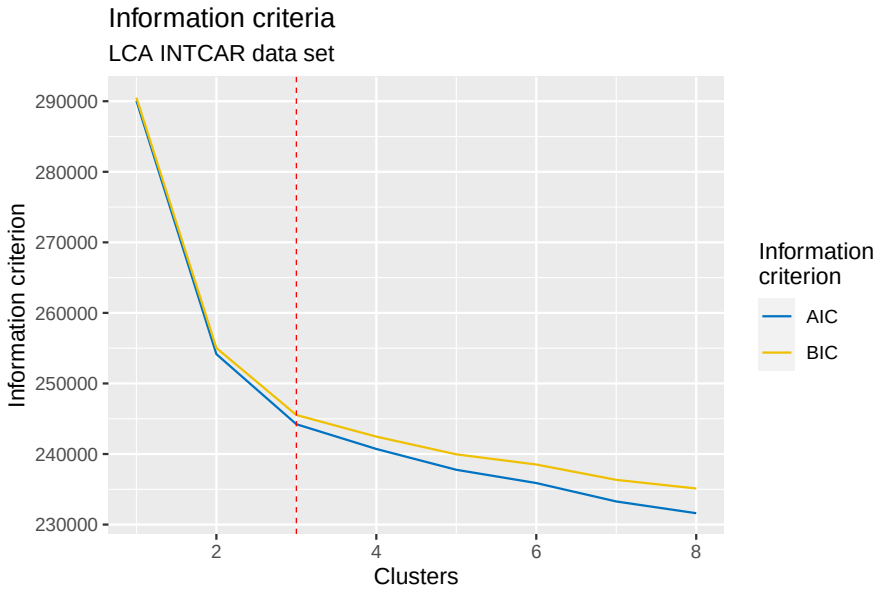

Figure 9: Plotting the AIC and BIC obtained when applying LCA to the INTCAR data set.

Table 19: CPC score and mortality by cluster as determined by LCA. Fisher's exact test, with simulated p-value where needed, is used to determine significance. How many patients that died due to withdrawal of life-support (WLS) are also stated.

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	***p*-value**
180-day CPC score, median [IQR]	5 [5, 5]	2 [1, 5]	2 [1, 5]	$< 0.001^*$
Mortality, n (%)	1619 (81)	554 (37)	268 (34)	$< 0.001^*$
WLS, n (%)	1315 ()	425 ()	177 ()	–

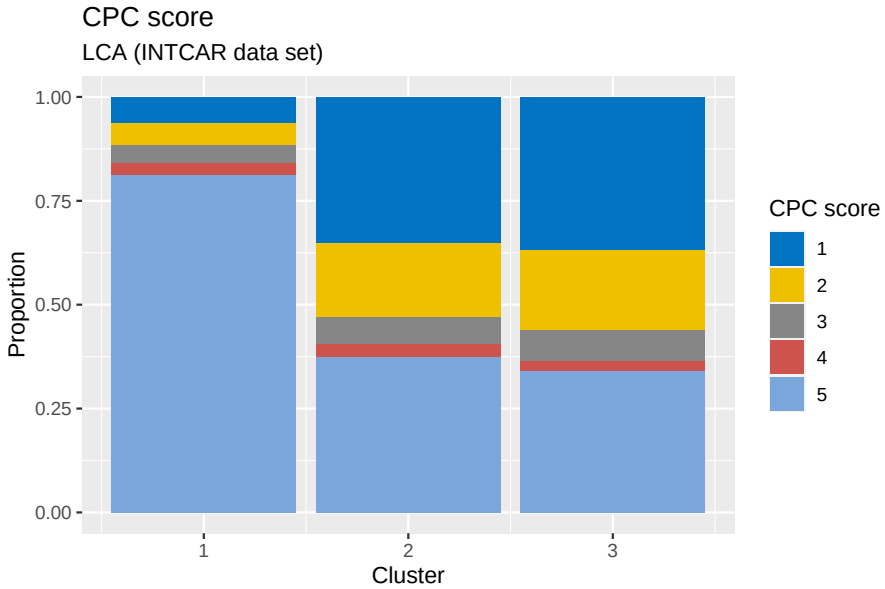

Figure 10: Distribution of the CPC score over different clusters obtained by LCA for the INTCAR data set.

4.2.2 KAMILA

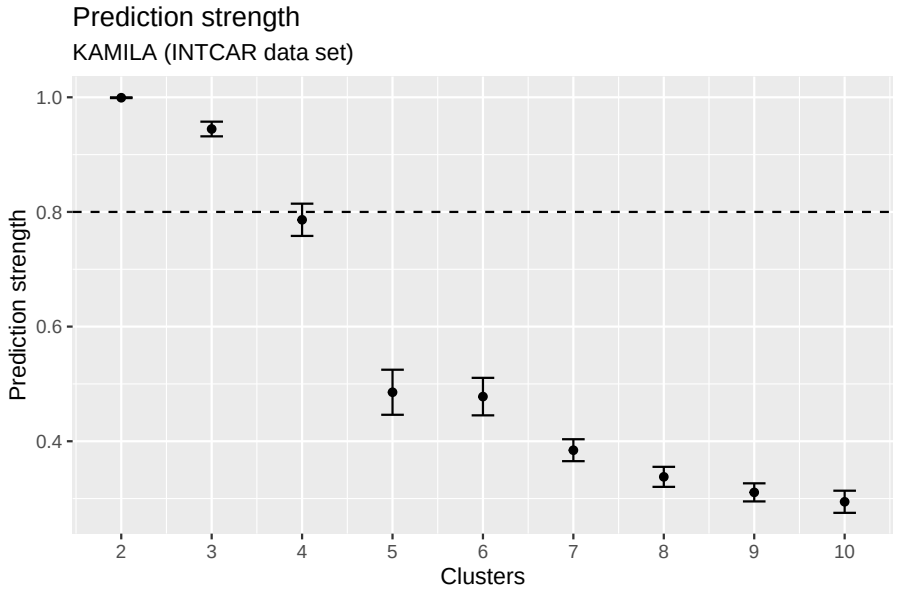

Figure 11: Prediction strength of the INTCAR data set for up to ten clusters where the error bars mark the standard errors. The dotted line marks the threshold of 0.8.

When applying KAMILA to the INTCAR data set, three clusters seem to be optimal according to prediction strength as plotted in Figure 11. This leads to the distribution of patients across the three clusters as stated in Table 20.

Table 20: The number of patients per cluster for the optimal solution when applying KAMILA to the INTCAR data set. The percentage is given in parentheses.

Cluster 1	**Cluster 2**	**Cluster 3**
1909 (45)	821 (19)	1531 (36)

In Figure 12 and Table 21 the differences in CPC score as well as mortality are shown. While the second and third cluster seem quite similar, especially when looking at the CPC score, the difference to the first cluster is very apparent. Here, less than 20% have a good neurological outcome, i.e. a CPC score of 1-3, and mortality is over 80%, which is more than double what it is in the other two clusters.

Table 21: CPC score and mortality by cluster as determined by KAMILA. Fisher's exact test, with simulated p-value where needed, is used to determine significance.

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	***p*-value**
180-day CPC score, median [IQR]	5 [5, 5]	2 [1, 5]	2 [1, 5]	$< 0.001^*$
Mortality, n (%)	1570 (82)	275 (33)	596 (39)	$< 0.001^*$
WLS, n (%)	1272 (81)	189 (69)	456 (77)	–

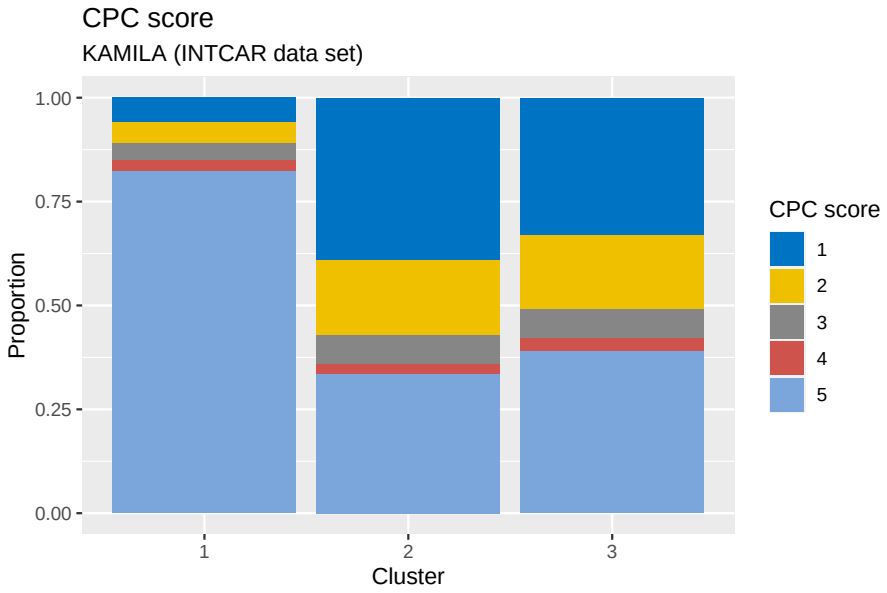

Figure 12: Distribution of the CPC score over different clusters obtained by KAMILA for the INTCAR data set.

4.2.3 k-prototypes

Since the average silhouette width reaches its maximum for two clusters as shown in Figure 13, the data is clustered into two clusters which are approximately of the same size, see Table 22.

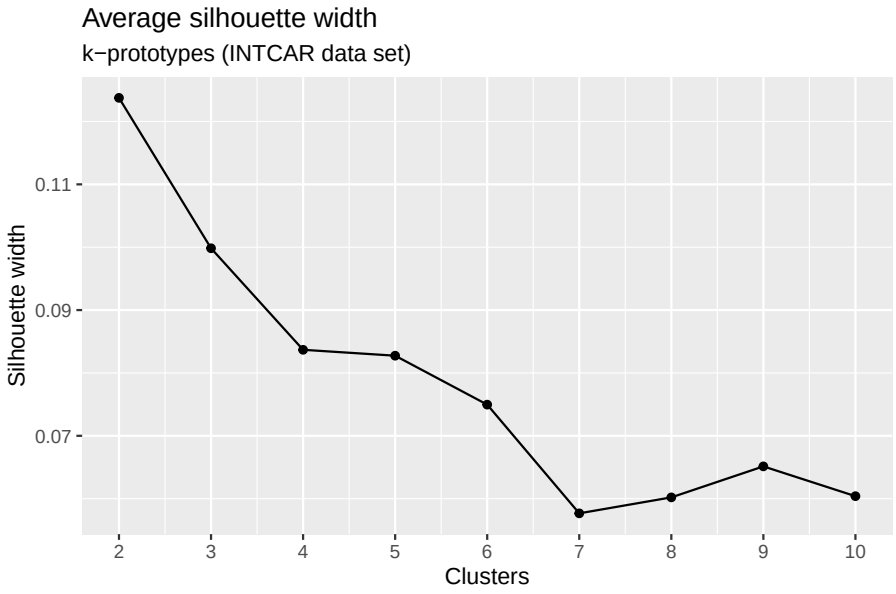

Figure 13: Average silhouette width when applying k-prototypes to the INTCAR data set.

Table 22: The number of patients per cluster for the optimal solution when applying k-prototypes to the INTCAR data set. The percentage is given in parentheses.

Cluster 1	**Cluster 2**
2242 (53)	2019 (47)

Even though the two clusters are similar in size, the outcome as measured by CPC score and mortality varies a lot across these two groups and with a p-value of less than 0.05, this difference is deemed significant. The first cluster has, not only, on average a much lower CPC score, but the mortality is also almost half of what it is in the second cluster. These results are presented in Figure 14 as well as Table 23.

Table 23: CPC score and mortality by cluster as determined by k-prototypes. Fisher's exact test, with simulated p-value where needed, is used to determine significance. How many patients that died due to withdrawal of life-support (WLS) are also stated.

Variable	**Cluster 1**	**Cluster 2**	***p*-value**
180-day CPC score, median [IQR]	2 [1, 5]	5 [5, 5]	$< 0.001^*$
Mortality, n (%)	868 (39)	1573 (78)	$< 0.001^*$
WLS, n (%)	625 (72)	1292 (82)	–

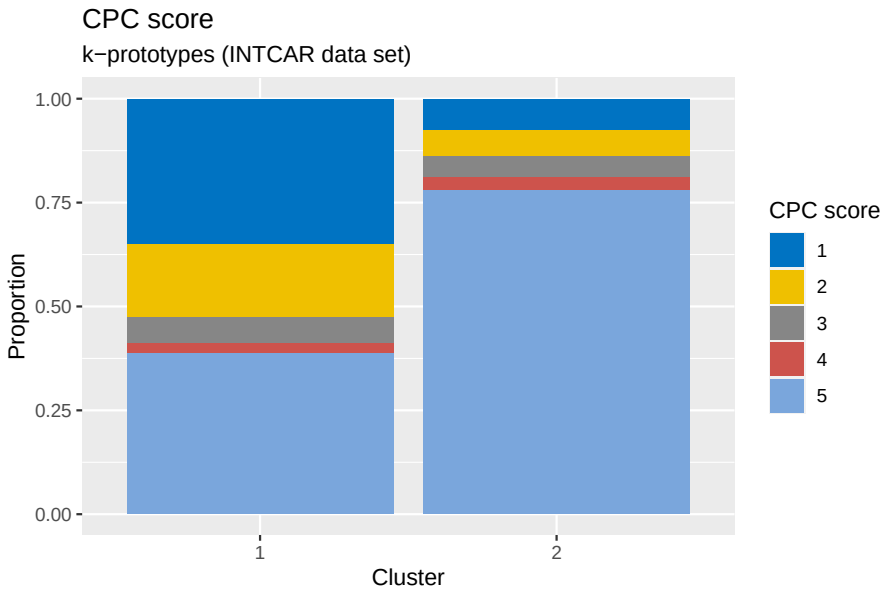

Figure 14: Distribution of the CPC score over different clusters obtained by k-prototypes for the INTCAR data set.

4.2.4 PAM

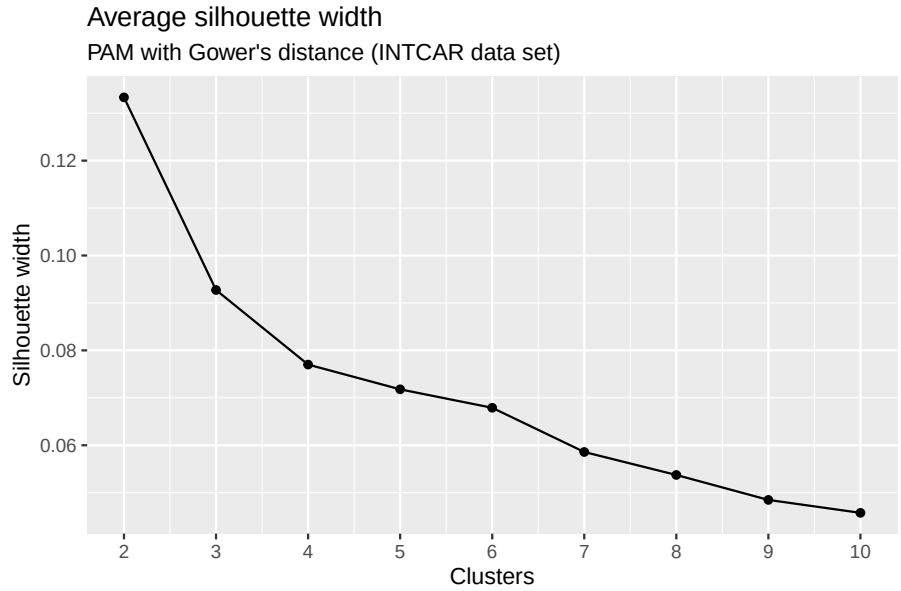

Figure 15: Average silhouette width when applying PAM with Gower's distance to the INTCAR data set.

As for k-prototypes, the average silhouette width reaches its maximum for two clusters when applying PAM with Gower's distance to the INTCAR data set, as shown in Figure 15.

In Table 24 it can be seen that the number of patients per cluster is, once again, quite evenly distributed. However, the outcomes are quite different for the two clusters. The distribution of the CPC score across the two different clusters is shown in Figure 16. Additional information not only about the CPC score but also about mortality is found in Table 25.

Table 24: The number of patients per cluster for the optimal solution when applying PAM with Gower's distance to the INTCAR data set. The percentage is given in parentheses.

Cluster 1	**Cluster 2**
2277 (53)	1984 (47)

Table 25: CPC score and mortality by cluster as determined by PAM. Fisher's exact test, with simulated p-value where needed, is used to determine significance. How many patients that died due to withdrawal of life-support (WLS) are also stated.

Variable	**Cluster 1**	**Cluster 2**	***p*-value**
180-day CPC score, median [IQR]	5 [5, 5]	4 [1, 5]	$< 0.001^*$
Mortality, n (%)	908 (40)	1533 (77)	$< 0.001^*$
WLS, n (%)	651 (72)	1266 (83)	–

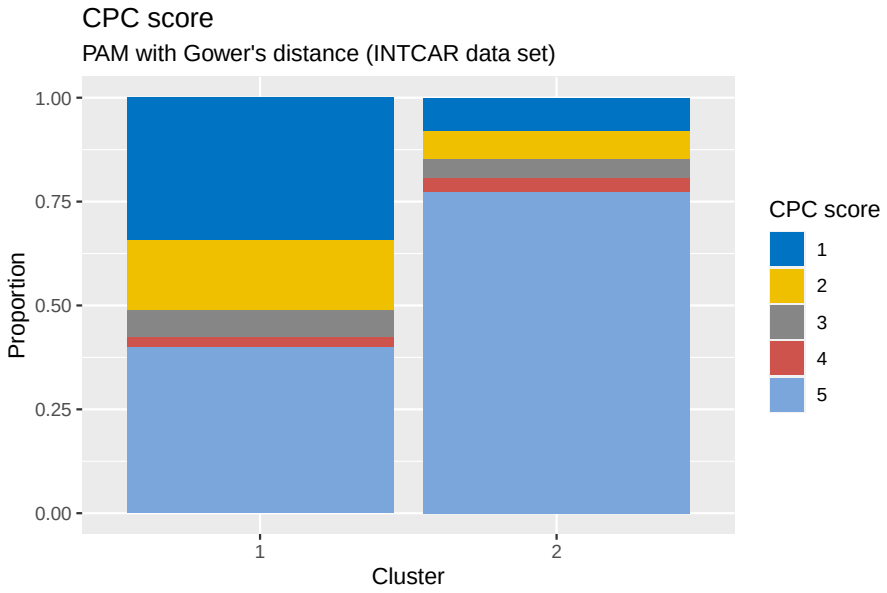

Figure 16: Distribution of the CPC score over different clusters obtained by PAM with Gower's distance for the INTCAR data set.

4.2.5 Logistic regression

As for the first data set, we now want to predict the binary CPC score of each patient. Once again, this is done with three different logistic regression models, the reference model, a model containing the cluster assignment as well as the TTM variable and lastly, a combination of these two models. Those results are presented in Tables 26 and 27. As for the first data set, the highest AUC-value is obtained when using the clustering assignment as given by KAMILA. When only using cluster assignment and whether targeted temperature management has been applied, we get an AUC-value of 0.7457. When combining those with the reference variables, we even get an AUC-value of 0.8626, which is also significantly better than when only using the reference variables. Similar to the results from the first data set, when using the clusters as given by LCA, we get the second-highest AUC values. Once again, the difference between using the cluster assignment as a factor variable or the cluster probabilities is not big. It is surprising that k-prototypes and PAM with Gower's distance perform much better for this data set than for the first one, with the AUC being not much smaller than for the best performing model.

Table 26: Comparing the area under the curve (AUC) of the reference model to the clustering as well as to the combined model where we try to predict the binary CPC. Here we test whether the AUC of the reference model is smaller than the AUC of the other model.

	LCA (Cluster membership)			**LCA (Cluster probabilities)**		
Model	**AUC**	**95% CI**	***p*-value**	**AUC**	**95% CI**	***p*-value**
Reference	0.8433	$(0.831 - 0.855)$	–	0.8433	$(0.831 - 0.855)$	–
Clustering	0.7291	$(0.714 - 0.744)$	1	0.7451	$(0.730 - 0.760)$	1
Combined	0.8589	$(0.847 - 0.871)$	$< 0.001^*$	0.8593	$(0.848 - 0.871)$	$< 0.001^*$

Thus, the cluster assignment as obtained by KAMILA leads to the best results when trying to predict the binary CPC score. These ROC curves are shown in Figure 17. Furthermore, a list of the included variables, as well as their odds ratio and corresponding confidence interval for both the reference and the clustering model, are given in Table 28. Here, it should be noted that not only the cluster assignment variables are significant but also the interaction between the second cluster and TTM.

Table 27: Comparing the area under the curve (AUC) of the reference model to the clustering as well as to the combined model where we try to predict the binary CPC score. Here we test whether the AUC of the reference model is smaller than the AUC of the other model.

	KAMILA			*k*-prototypes			PAM		
Model	**AUC**	**95% CI**	***p*-value**	**AUC**	**95% CI**	***p*-value**	**AUC**	**95% CI**	***p*-value**
Reference	0.8433	(0.831 – 0.855)	–	0.8434	(0.831 – 0.856)	–	0.8434	(0.831 – 0.856)	–
Clustering	0.7457	(0.732 – 0.760)	1	0.7168	(0.703 – 0.731)	1	0.7039	(0.690 – 0.718)	1
Combined	0.8626	(0.851 – 0.874)	$< 0.001^*$	0.8577	(0.846 – 0.869)	$< 0.001^*$	0.8551	(0.844 – 0.867)	$< 0.001^*$

Table 28: Odds ratio (OR) and the corresponding confidence interval (CI) of the reference model as well as of the clustering models obtained by KAMILA using the cluster assignment as factor variable.

	Reference Model			Clustering Model KAMILA		
Variable	**OR**	**95% CI**	***p*-value**	**OR**	**95% CI**	***p*-value**
Intercept	17.77	(13.73 – 23.18)	$< 0.001^*$	12.29	(7.74 – 20.93)	$< 0.001^*$
Age	1.82	(1.68 – 1.98)	$< 0.001^*$	–	–	–
Sex	0.65	(0.55 – 0.78)	$< 0.001^*$	–	–	–
Bystander witnessed arrest	0.47	(0.37 – 0.58)	$< 0.001^*$	–	–	–
First rhythm shockable	0.16	(0.13 – 0.19)	$< 0.001^*$	–	–	–
ROSC	2.20	(2.01 – 2.41)	$< 0.001^*$	–	–	–
GCS Motor	0.64	(0.60 – 0.70)	$< 0.001^*$	–	–	–
Cluster 2	–	–	–	0.04	(0.02 – 0.07)	0.003^*
Cluster 3	–	–	–	0.08	(0.04 – 0.15)	$< 0.001^*$
TTM	–	–	–	0.64	(0.37 – 1.04)	0.087
Cluster 2 · TTM	–	–	–	2.79	(1.49 – 5.47)	0.002^*
Cluster 3 · TTM	–	–	–	1.61	(0.79 – 3.41)	0.201

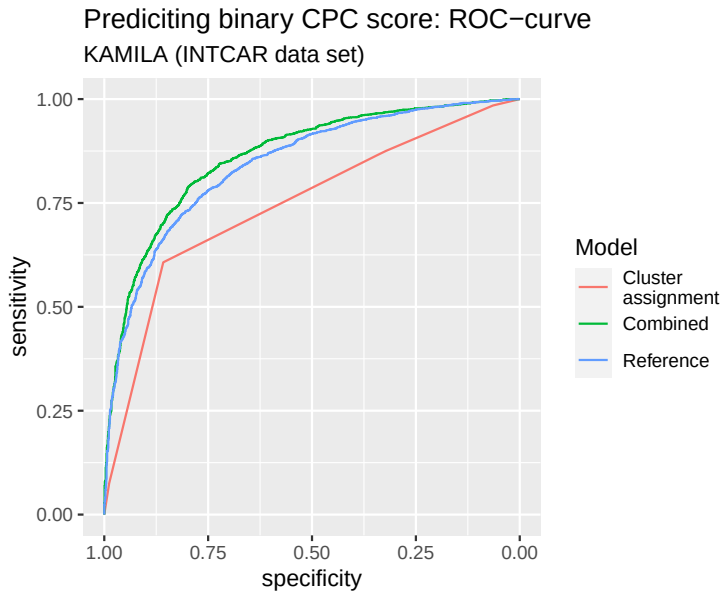

Figure 17: ROC curves of the three different logistic regression models where the cluster assignment is obtained by applying KAMILA to the INTCAR data set.

5 Discussion

When trying to cluster cardiac arrest patients into distinct subgroups, different methods lead to different results, so it is difficult to conclude how many subgroups there truly are. For the TTM data set, LCA indicates that are six clusters, whereas the other three models indicate that there are only two. Furthermore, even among the three methods that indicate two clusters, the number of patients and distributions of both clinical and outcome variables vary between methods. While KAMILA and LCA indicate that there are three clusters in the second data set, k-prototypes and PAM with Gower's distance indicate that there are only two clusters. Once again there are differences even amongst the methods that indicate the same number of clusters.

Comparing the clusters of the two different data sets is complicated even further since the number of variables contained in both data sets is very small. A summary over which variables are contained in both data sets is given in Table 29. Furthermore, this table also includes five additional variables that are part of both data sets but are not of the same format. For example, the GCS Motor variable has categories 1-6 in the INTCAR data set but has an additional category, 9, in the TTM data set. Another example is the variable Seizures. In the INTCAR data set, it is a binary variable denoting whether or not the patient has had any seizures, whereas in the TTM data set it measures how many seizures a patient has had before admission.

Table 29: Variables that are part of both data sets are above the bold line. For the continuous variables the mean and the standard deviation are reported. For the categorical variables the number of observations in each category and their corresponding percentages are reported. For the binary variables the number of observations in the affirmative category and their corresponding percentages are reported. All variables below the double line are part of both data sets but on different scales.

Variable	**TTM data set**	**INTCAR data set**
Age, years	64.1 ± 12.2	61.3 ± 15.5
ROSC, minutes	31.4 ± 22.3	26.1 ± 17.7
Temperature, °C	35.26 ± 1.18	35.34 ± 1.47
Compression		
No	714 (77)	2918 (68)
Yes, manual	65 (7)	371 (9)
Yes, mechanic	154 (16)	864 (20)
Sex		
Male	758 (81)	2926 (69)
Witnessed cardiac arrest	835 (89)	3350 (79)
Bystander CPR	681 (73)	2789 (65)
Shock on admission	137 (15)	1572 (37)
Arterial hypertension	373 (39.9)	1964 (46.1)
GCS Motor (1-6)	1.62 ± 1.18	1.49 ± 1.14
Diabetes	140 (15)	945 (22)
Malignancy	31 (3)	243 (6)
Defibrillations	825 (88)	2727 (64)
Seizures	58 (6)	670 (16)

There are also some issues with the clustering methods, which can possibly explain the different results. One of the key assumptions of LCA is that the continuous variables follow a Gaussian distribution. While transformations are applied to reduce skewness, even after those, not all continuous variables seem to follow a Gaussian distribution. Additionally, it is also assumed that variables within different clusters are locally independent, and this assumption does not hold. Furthermore, when applying LCA to the INTCAR data set, the BIC does not converge, and hence cluster enumeration is based on the elbow criterion. While this usually is a valid alternative, cluster enumeration for the two different data sets is thus not based on exactly the same criterion.

When using KAMILA, the number of clusters k is chosen as the maximum k such that $ps(k) > 0.8$ since it has been shown that a threshold of 0.8 leads to good clustering results for well-separated clusters. However, it is not known whether or not the clusters are well separated. Furthermore, even if we know that they are not well separated, there seems to be no consensus on which threshold to use in that case.

Similarly, when using k-prototypes, one needs to estimate the weighting parameter λ. While the choice of λ does not seem to affect the optimal number of clusters for our two data sets, it still affects the final clustering results. Even if the optimal number of clusters is $k = 2$ for different λ, different λ still lead to different clusters.

However, in spite of all the problems described above, the KAMILA algorithm leads to the best results when predicting the binary CPC score for both data sets. For both data sets, an AUC-value of more than 0.69 indicates quite good classification results since a value greater than 0.5 implies better than random classification. Those results are promising since this is achieved only by including the cluster assignment variables as well as the temperature management variable. Additionally, in the TTM data set, the cluster assignment variable has a p-value of less than 0.05 and thus is statistically significant. In the INTCAR data set, both the cluster assignment variable as well as the interaction variable between being in the second cluster and TTM being used is also statistically significant.

Even the results from the combined models are promising. If the variables from the reference model, i.e., variables that are deemed significant from a medical perspective, are also included, we obtain an AUC-value of more than 0.8 for both data sets. Furthermore, the AUC-value of the combined model is also greater than the AUC-value of the reference model with a statistically significant p-value. This implies that adding both the cluster assignment and TTM variable adds new, relevant information to the model.

Furthermore, when looking at how the variables are distributed among the different clusters, see Appendix A for results for the TTM data set and Appendix B for the results from the INTCAR data set, the results make sense even from a medical point of view. For example, in the TTM data set, patients in the first cluster obtained by KAMILA have a mortality rate of more than twice that of patients in the second cluster. This can (possibly) be explained by them on average having a lower GCS Motor score, higher lactate values and longer time between cardiac arrest and return of spontaneous circulation (ROSC) than patients from the second cluster. Similarly, when analysing the clustering obtained by KAMILA when applied to the data from the second data set, the patients in the second cluster have on average the best neurological outcome, i.e. a low CPC score, and the lowest mortality. When looking at the clinical variables, those patients also have a high GCS Motor score, little time between cardiac arrest and return of spontaneous circulation, as well as the majority breathing when admitted to the hospital.

6 Conclusion

In conclusion, there seems to be some evidence that cardiac arrest patients can be divided into different clusters with different underlying distributions. Furthermore, for almost all methods, there is a significant difference in outcome for patients in different clusters, with the exception being PAM with Gower's distance when it is applied to the TTM data set. The outcome is measured by the CPC score as well as mortality. When predicting the binary CPC score, all methods achieve an AUC-value greater than 0.5, which means that classification is better than random. The most promising results are obtained by the KAMILA algorithm with an AUC-value of 0.6965 for the TTM data set and 0.7457 for the INTCAR data set. In both cases, only the targeted temperature management and cluster assignment variables are used to obtain those AUC-values. Including medically relevant variables improves the results even further.

The results from the INTCAR data set also indicate that the effect of cooling down patients might vary across different clusters and might be more useful to some clusters than others. More precisely, while the targeted temperature management variable on its own is not statistically significant, the interaction with being in the second cluster is statistically significant. However, it is still unclear why the patients in the second cluster seem to benefit from being cooled down.

Additionally, further research needs to be done to find out which clinical variables are most important when clustering cardiac arrest patients into different subgroups. This can possibly lead to improved clustering and prediction results.

7 References

[1] *About Cardiac Arrest*. American Heart Association. URL: https://www.heart.org/en/health-topics/cardiac-arrest/about-cardiac-arrest. (accessed: 06.05.2021).

[2] *Cardiac Arrest*. British Heart Foundation. URL: https://www.bhf.org.uk/informationsupport/conditions/cardiac-arrest. (accessed: 06.05.2021).

[3] Daniel B. Knox et al. "Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome". In: *Intensive Care Medicine* 41.5 (2015), pp. 814–822. DOI: https://doi.org/10.1007/s00134-015-3764-7.

[4] Bengt Gårdlund et al. "Six subphenotypes in septic shock: Latent class analysis of the PROWESS Shock study". In: *Journal of Critical Care* 47 (2018), pp. 70–79. DOI: https://doi.org/10.1016/j.jcrc.2018.06.012.

[5] Amir Ahmad and Shehroz Khan. "Survey of State-of-the-Art Mixed Data Clustering Algorithms". In: *IEEE Access* 7 (2019), pp. 31883–31902. DOI: https://doi.org/10.1109/ACCESS.2019.2903568.

[6] Erich Schubert and Peter J. Rousseeuw. *Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms*. Springer International Publishing, 2019, pp. 171–187. DOI: https://doi.org/10.1007/978-3-030-32047-8_16.

[7] Zhexue Huang. "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values." In: *Data Mining and Knowledge Discovery* 2.3 (1998), pp. 283–304. DOI: https://doi.org/10.1023/A:1009769707641.

[8] Alex Foss and Marianthi Markatou. "kamila: Clustering Mixed-Type Data in R and Hadoop". In: *Journal of Statistical Software* 83.13 (2018), pp. 1–44. DOI: https://doi.org.10.18637/jss.v083.i13.

[9] Alice Kongsted and Anne Molgaard Nielsen. "Latent Class Analysis in health research". In: *Journal of Physiotherapy* 63.1 (2017), pp. 55–58.

[10] Karen Nylund-Gibson and Andrew Young Choi. "Ten Frequently Asked Questions About Latent Class Analysis". In: *American Psychological Association* 4.4 (2018), pp. 440–461. DOI: https://doi.apa.org/doiLanding?doi=10.10372Ftps0000176.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, 2017. ISBN: 978-0-387-84857-0.

[12] Jay Magidson and Jeroen K. Vermunt. *Latent Class Models*. Ed. by David Kaplan. SAGE Publications, Inc., 2004. Chap. 10, pp. 176–199. DOI: https://dx.doi.org/10.4135/9781412986311.

[13] Karen L. Nylund, Tihomir Asparouhov, and Bengt O. Muthén. "Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study". In: *Structural Equation Modeling: A Multidisciplinary Journal* 14.4 (2007), pp. 535–569. DOI: https://doi.org/10.1080/10705510701575396.

[14] Jeroen K. Vermunt and Jay Magidson. *Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc., 2016.

[15] Alex Foss et al. "A semiparametric method for clustering mixed data". In: *Machine Learning* 105 (2016), pp. 419–458. DOI: https://doi.org/0.1007/s10994-016-5575-7.

[16] Robert Tibshirani and Guenther Walther. "Cluster Validation By Prediction Strength". In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 511–528. DOI: https://doi.org/10.1198/106186005X59243.

[17] Gero Szepannek. "ClustMixType: User-Friendly Clustering of Mixed-Type Data in R." In: *R Journal* 10.2 (2018), pp. 200–208. ISSN: 20734859.

[18] J. C. Gower. "A General Coefficient of Similarity and Some of Its Properties". In: *Biometrics* 27.4 (1971), pp. 857–871.

[19] A. P. Reynolds et al. “Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms”. In: *Journal of Mathematical Modelling and Algorithms* 5.4 (2006), pp. 475–504. DOI: https://doi.org/10.1007/s10852-005-9022-1.

[20] Hae-Young Kim. “Statistical notes for clinical researchers: Chi-squared test and Fisher’s exact test”. In: *Restorative Dentistry Endodontics* 42.2 (2017), pp. 152–155. DOI: https://doi.org/10.5395%2Frde.2017.42.2.152.

[21] Julien I.E.Hoffman. *Hypergeometric Distribution*. Academic Press, 2015. Chap. 13, pp. 179–182. DOI: https://doi.org/10.1016/B978-0-12-802387-7.00013-5.

[22] Dao Lam, Mingzhen Wei, and Donald Wunsch. “Clustering Data of Mixed Categorical and Numerical Type With Unsupervised Feature Learning”. In: *IEEE Access* 3 (2015), pp. 1605–1613. DOI: https://doi.org/10.1109/ACCESS.2015.2477216.

[23] Magdalena Szumilas. “Explaining Odds Ratios”. In: *Journal of the Canadian Academy of Child Adolescent Psychiatry* 19.3 (2010), pp. 227–229. ISSN: 17198429.

[24] Kentaro Hayashi, Peter M.Bentler, and Ke-Hai Yuan. *Structural Equation Modeling*. Elsevier Science, 2011. Chap. 7, pp. 202–234. DOI: https://doi.org/10.1016/B978-0-444-53737-9.50010-4.

A Appendix: TTM data set

A.1 Included variables

Table 30: All variables from the TTM data set that are used to find the clusters. For the continuous variables the mean and the standard deviation are reported. For the binary and categorical variables the number of positives and their corresponding percentage is reported.

Variable	**Overall**
Age, years	64.1 ± 12.2
Weight, kg	82.3 ± 15.9
Defibrillations	3.55 ± 2.90
Seizures	0.09 ± 0.39
Cardiac arrest to advanced life support (ALS), min	10.08 ± 6.58
Cardiac arrest to return of spontaneous circulation (ROSC), min	31.4 ± 22.3
Adrenaline	3.62 ± 2.84
Temperature, °C	35.3 ± 1.2
GCS Motor	3.76 ± 3.49
B-glucose	13.86 ± 5.33
pO_2	24.4 ± 16.2
pCO_2	6.52 ± 2.19
Base excess	-8.62 ± 6.80
Potassium	3.97 ± 0.81
FiO_2	76.8 ± 23.4
Creatinine	114.3 ± 63.2
Platelets	229.1 ± 77.2
White blood cell count (WBC)	14.80 ± 6.01
pH	7.20 ± 0.16
Lactate	6.70 ± 4.46
Cardiac arrest location	
Place of residence	499 (53.4)
Public place	383 (41.0)
Other	53 (5.7)
First monitored rhythm	
Non-perfusing VT	24 (2.6)
VF	703 (75.2)
Asystole	113 (12.1)
PEA	65 (7.0)
Unknown	18 (1.9)
ROSC after bystander defibrillation	12 (1.3)
Automatic compression-decompression	
No	714 (76.4)
Yes, manual	65 (7.0)
Yes, mechanic	154 (16.5)
Sex	
Male	758 (81.1)
Congestive heart failure (CHF)	61 (6.5)
Previous acute myocardial infarction (AMI)	193 (20.6)
Ischemic heart disease (IHD)	259 (27.7)
Previous arrhythmia	165 (17.6)
Previous cardiac arrest	21 (2.2)
Arterial hypertension	373 (39.9)
Transient ischaemic attack (TIA) or stroke	73 (7.8)
Epilepsy	17 (1.8)
Diabetes	140 (15.0)

Table 30: All variables from the TTM data set that are used to find the clusters. For the continuous variables the mean and the standard deviation are reported. For the binary and categorical variables the number of positives and their corresponding percentage is reported.

Variable	**Overall**
Insulin	42 (4.5)
Asthma or chronic obstructive pulmonary disease (COPD)	97 (10.4)
Chronic dialysis	6 (0.6)
Cirrhosis	3 (0.3)
Hematological malignancy	9 (1.0)
Other malignancy	23 (2.5)
AIDS	1 (0.1)
Alcoholism	37 (4.0)
IV drug abuse	5 (0.5)
Immunodeficiency	4 (0.4)
Previous percutaneous coronary intervention (PCI)	107 (11.4)
Previous coronary artery bypass grafting (CABG)	89 (9.5)
Previous valvular surgery	25 (2.7)
Implantable cardioverter-defibrillator (ICD)	6 (0.6)
Pacemaker	34 (3.6)
Bystander witnessed arrest	835 (89.3)
Bystander CPR	681 (72.8)
Bystander defibrillation	92 (9.8)
First rhythm shockable	749 (80.1)
Pre-hospital intubation	625 (66.8)
Acute ST-infarction or LBBB	437 (46.7)
Pupillary reflex	702 (75.1)
Corneal reflex	519 (55.5)
Cough reflex	490 (52.4)
Spontaneous breathing	598 (64.0)
Shock on admission	137 (14.7)
GCS motor 1	491 (52.5)
GCS motor 2	39 (4.2)
GCS motor 3	45 (4.8)
GCS motor 4	62 (6.6)
GCS motor 5	23 (2.5)
GCS motor 6	1 (0.1)
GCS motor 9	269 (28.8)

A.2 LCA

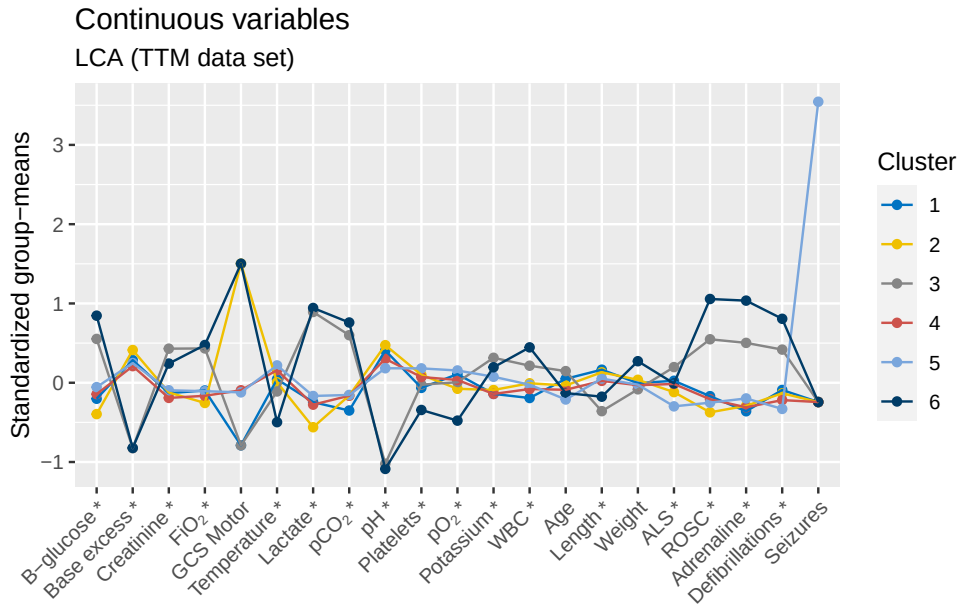

Figure 18: Distribution of the continuous variables of the TTM data set across different clusters as obtained by LCA. If there is a significant difference, the variable is marked by an asterisk, *. For the variables GCS Motor and Seizures it is not possible to determine significance with ANOVA.

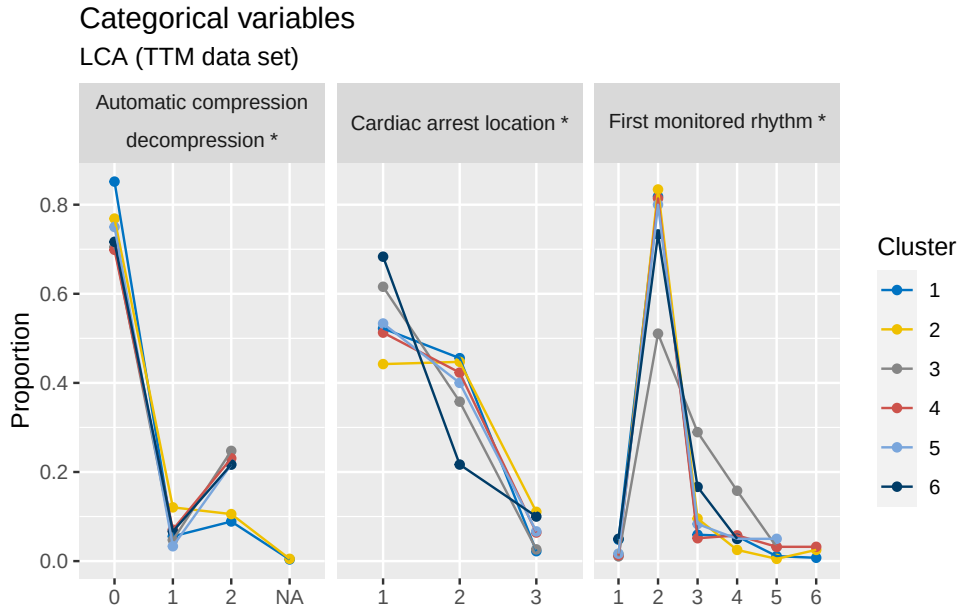

Figure 19: Distribution of the categorical variables of the TTM data set across different clusters as obtained by LCA. If there is a significant difference, the variable is marked by *.

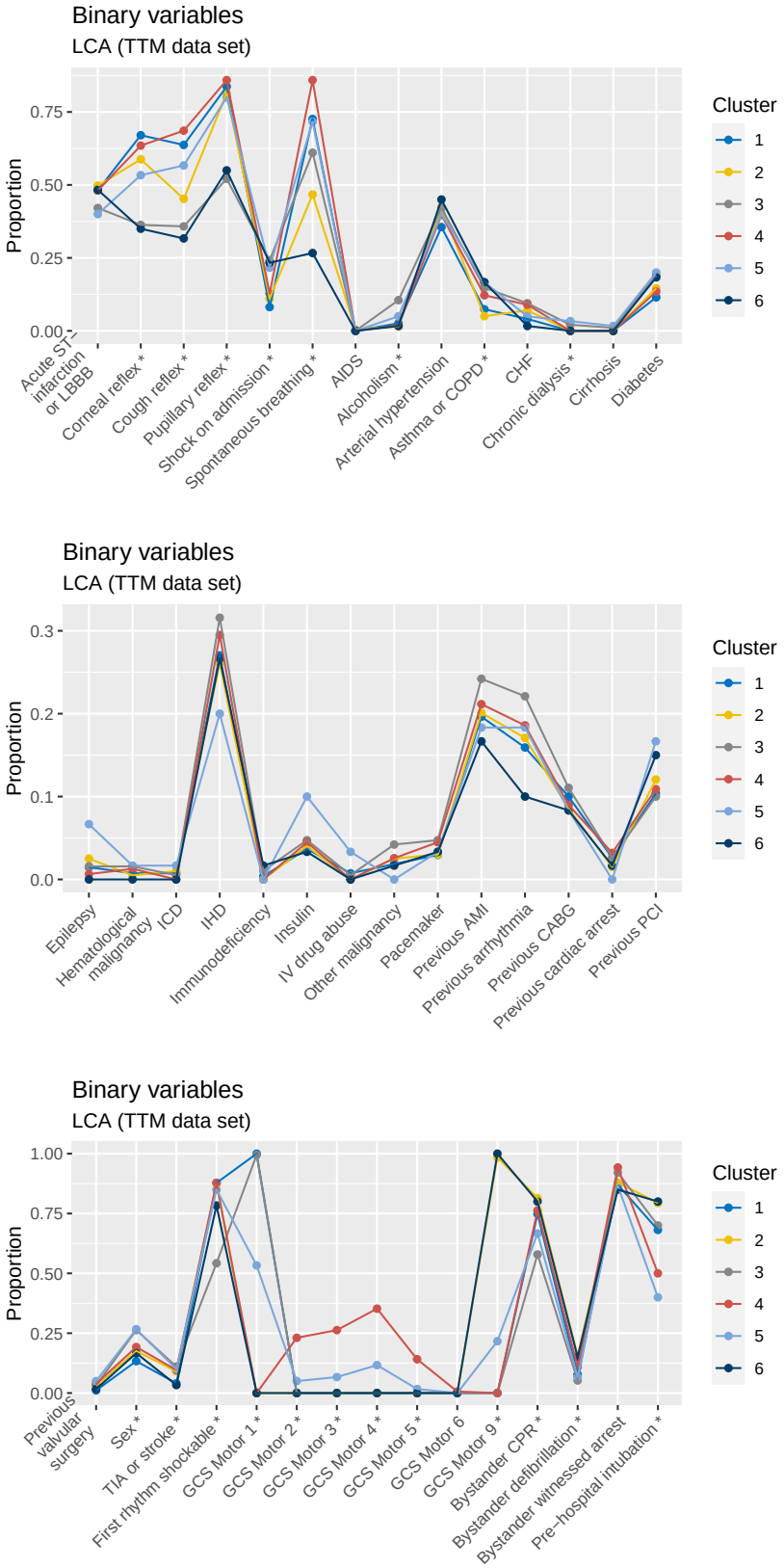

Figure 20: Distribution of the binary variables of the TTM data set across different clusters as obtained by LCA. If there is a significant difference, the variable is marked by an asterisk, *.

A.3 KAMILA

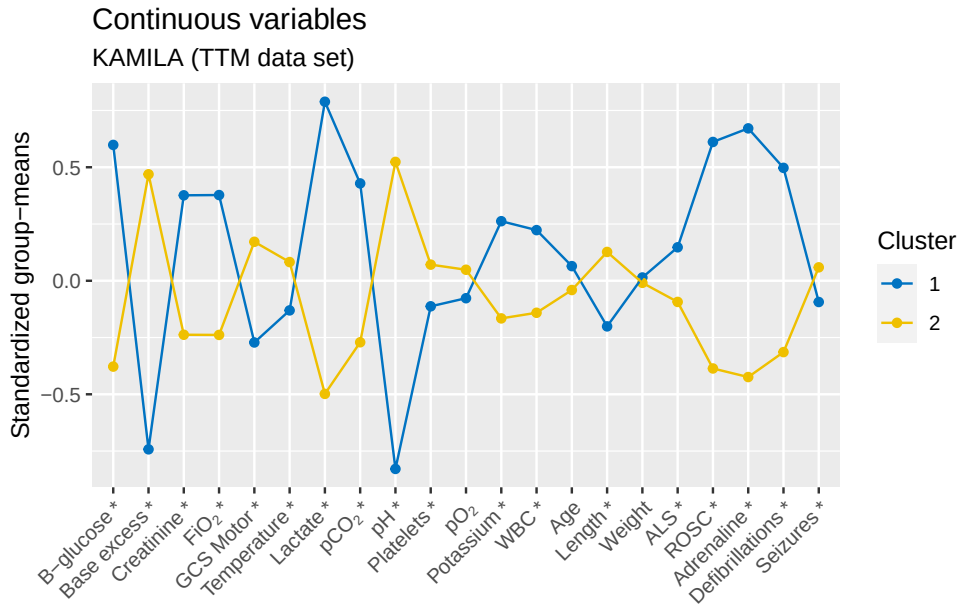

Figure 21: Distribution of the continuous variables of the TTM data set across different clusters as obtained by KAMILA. If there is a significant difference, the variable is marked by an asterisk, *.

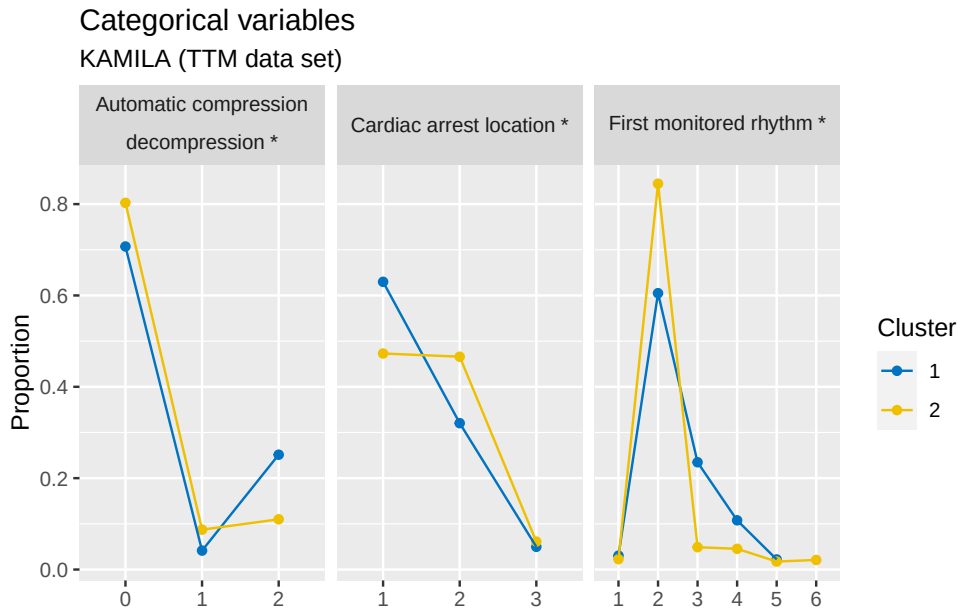

Figure 22: Distribution of the categorical variables of the TTM data set across different clusters as obtained by KAMILA. If there is a significant difference, the variable is marked by an asterisk, *.

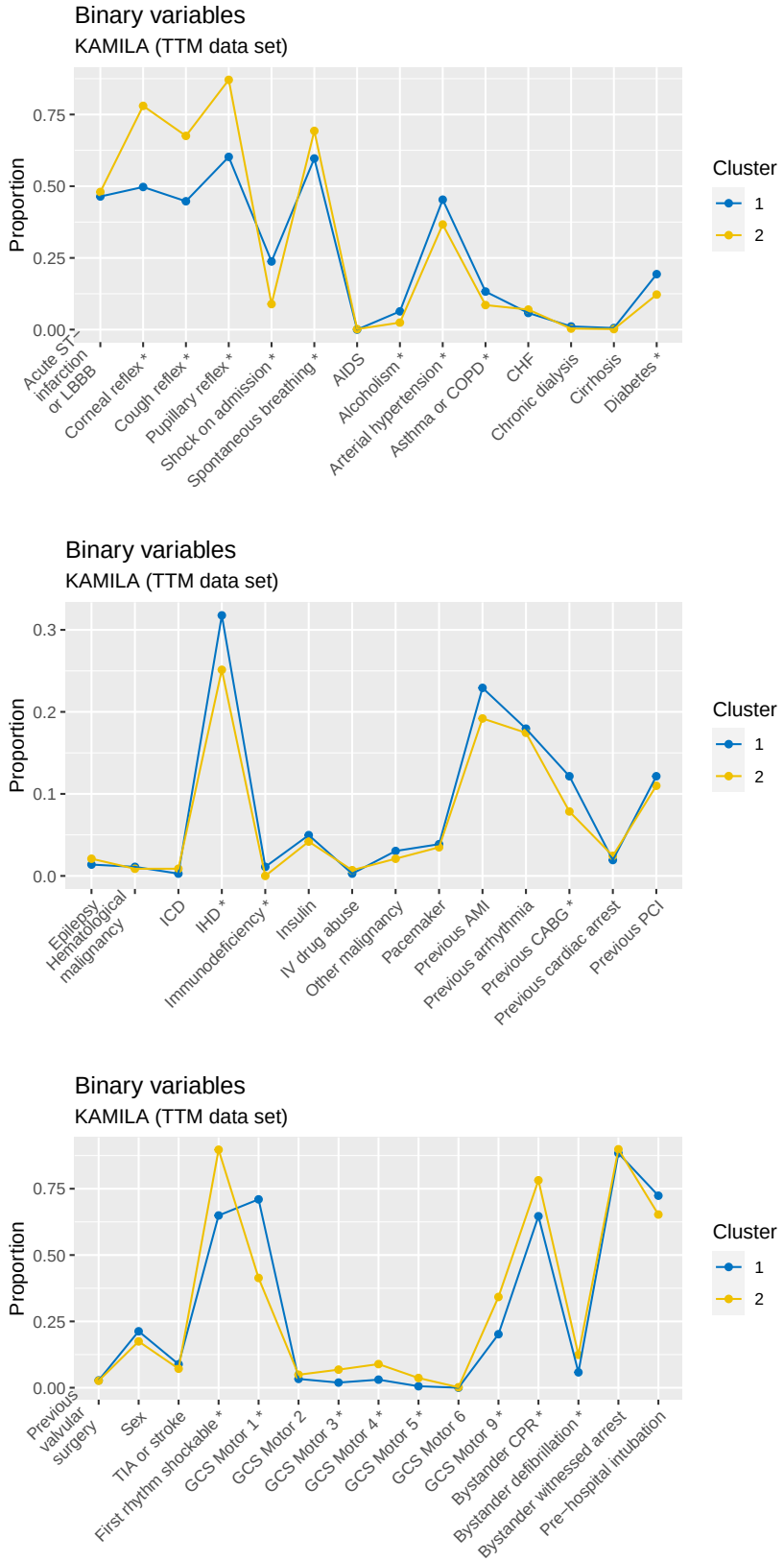

Figure 23: Distribution of the binary variables of the TTM data set across different clusters as obtained by KAMILA. If there is a significant difference, the variable is marked by an asterisk, *.

A.4 k-prototypes

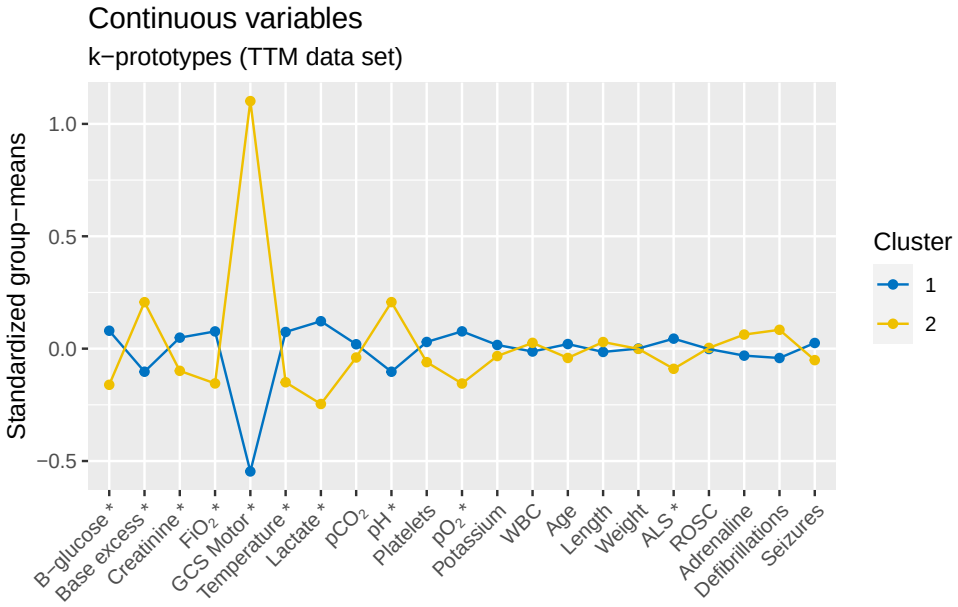

Figure 24: Distribution of the continuous variables of the TTM data set across different clusters as obtained by k-prototypes. If there is a significant difference, the variable is marked by an asterisk, *.

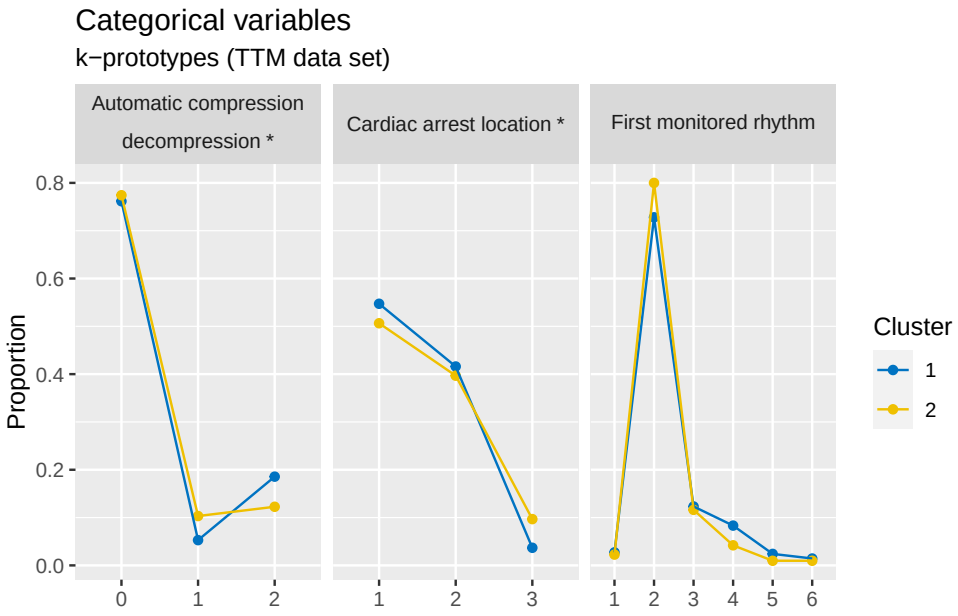

Figure 25: Distribution of the categorical variables of the TTM data set across different clusters as obtained by k-prototypes. If there is a significant difference, the variable is marked by an asterisk, *.

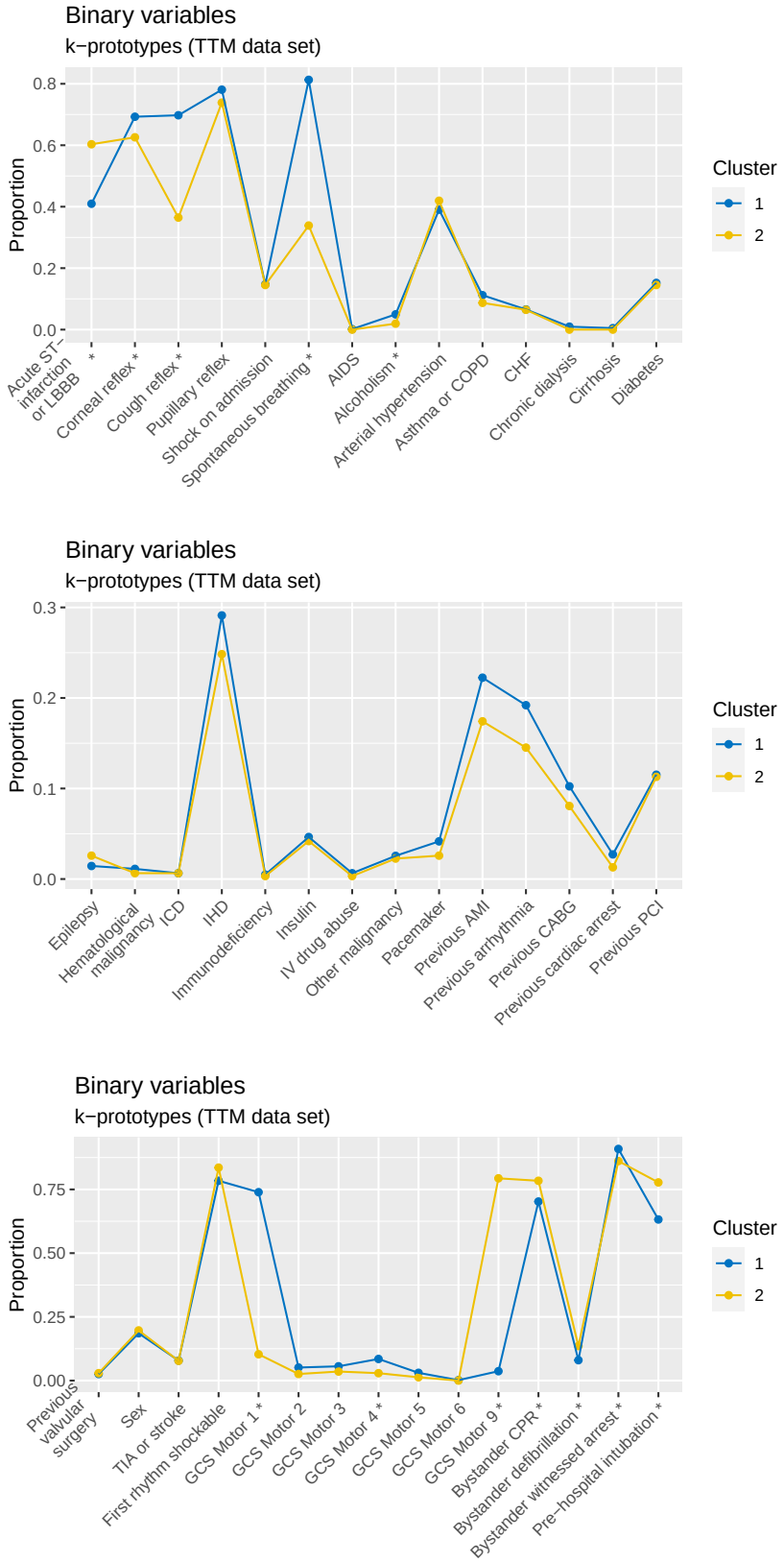

Figure 26: Distribution of the binary variables of the TTM data set across different clusters as obtained by k-prototypes. If there is a significant difference, the variable is marked by an asterisk, *.

A.5 PAM

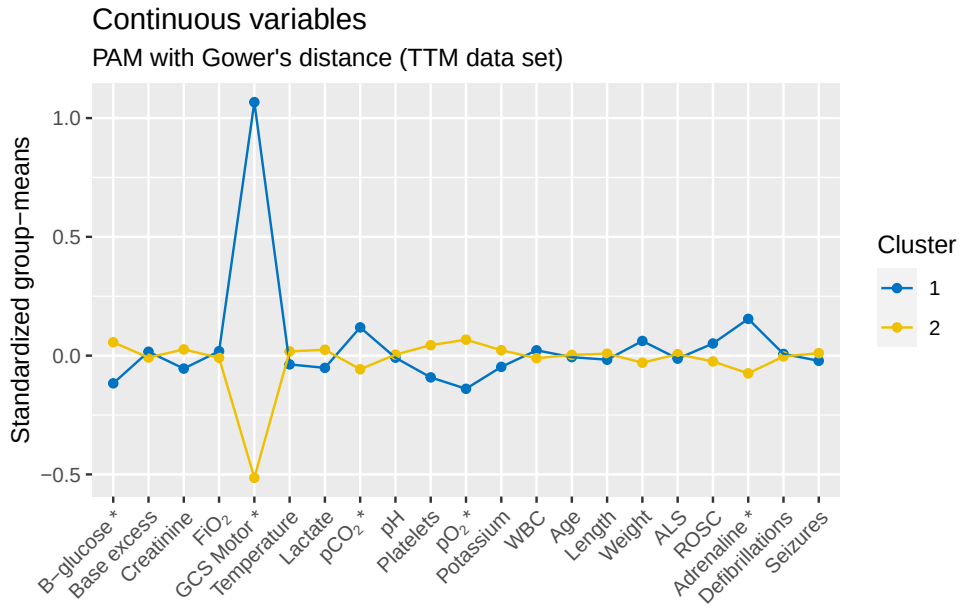

Figure 27: Distribution of the continuous variables of the TTM data set across different clusters as obtained by PAM with Gower's distance. If there is a significant difference, the variable is marked by an asterisk, *.

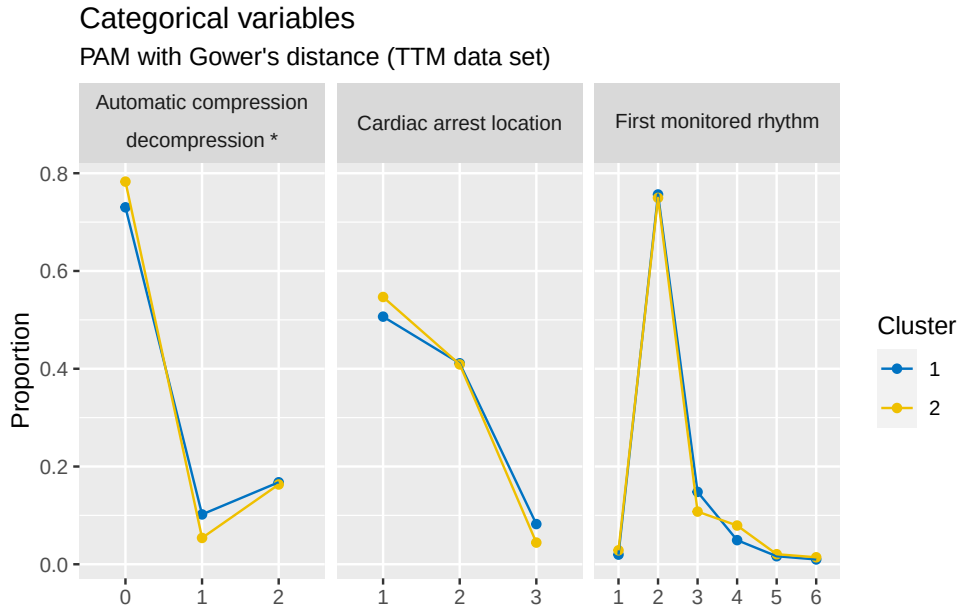

Figure 28: Distribution of the categorical variables of the TTM data set across different clusters as obtained by PAM with Gower's distance. If there is a significant difference, the variable is marked by an asterisk, *.

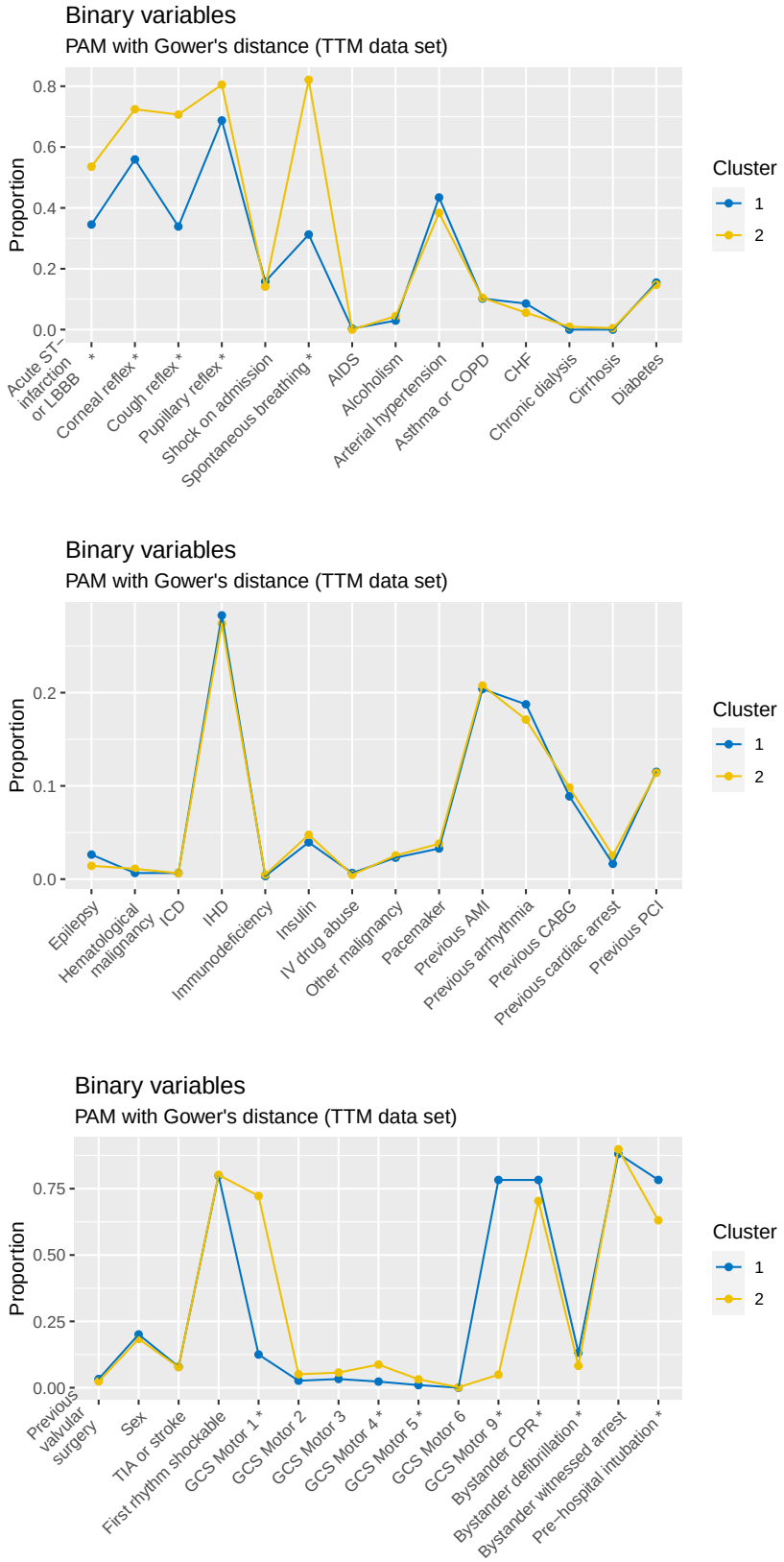

Figure 29: Distribution of the continuous variables of the TTM data set across different clusters as obtained by PAM with Gower's distance. If there is a significant difference, the variable is marked by an asterisk, *.

B Appendix: INTCAR data set

B.1 Included variables

Table 31: All variables from the INTCAR data set that are used to find the clusters. For the continuous variables the mean and the standard deviation are reported. For the binary and categorical variables the number of positive outcomes and their corresponding percentage is reported.

Variable	**Overall**
Age, years	61.3 ± 15.5
Time to emergency medical services (EMS) to CPR, min	8.42 ± 6.40
Cardiac arrest to return of spontaneous circulation (ROSC), min	26.1 ± 17.7
Temperature, °C	35.34 ± 1.47
GCS Motor	1.49 ± 1.14
Rhythm	
PEA/Asystole (unshockable)	1793 (42.1)
Unknown (unshockable)	181 (4.2
VT/VF (shockable)	2264 (53.1)
Compression	
Yes, manual	371 (8.7)
Yes, mechanic	864 (20.3
No	2918 (68.5)
Commands	
No	3271 (76.8)
Sedated	757 (17.8)
Yes	99 (2.3)
Breath	
Could not be determined	824 (19.3)
No	1805 (42.4)
Yes	1559 (36.6)
Awake at cath lab	
Awake	285 (6.7)
No cath lab	2049 (48.1)
Unconscious	1896 (44.5)
ECG	
Abnormal	2231 (52.4)
LBBB	272 (6.4)
Normal	561 (13.2)
No ECG	115 (2.7)
STEMI	1015 (23.8)
Echo	
$EF < 30$	855 (20.1)
$EF > 50$	1153 (27.1)
$EF \quad 30 - 49$	1089 (25.6)
None	942 (22.1)
Sex	
Male	2926 (69)
Pulmonary hypertension	813 (19.1)
Coronary artery disease	1059 (24.9)
Congestive heart failure	753 (17.7)
Arrhythmia	588 (13.8)
Chronic obstructive pulmonary disease (COPD)	636 (14.9)
Arterial hypertension	1964 (46.1)
Chronic kidney disease	406 (9.5)
Neurological disease	507 (11.9)

Table 31: All variables from the INTCAR data set that are used to find the clusters. For the continuous variables the mean and the standard deviation are reported. For the binary and categorical variables the number of positive outcomes and their corresponding percentage is reported.

Variable	**Overall**
Liver disease	108 (2.5)
Malignancy	243 (5.7)
Obesity	431 (10.1)
Insulin-dependent diabetes mellitus (IDDM)	384 (9.0)
Non-insulin-dependent diabetes mellitus (NIDDM)	571 (13.4)
Witnessed cardiac arrest	3350 (78.6)
Bystander CPR	2789 (65.5)
Defibrillations	2727 (64.0)
CT on admission	1921 (45.1)
Shock	1572 (36.9)
CT	2287 (53.7)
MRI	648 (15.2)
EEG	2288 (53.7)
cEEG	1660 (39)
Somatosensory evoked potential (SSEP)	500 (11.7)
Seizure	670 (15.7)
Myoclonus	1091 (25.6)
AEDs	467 (11.0)
Thrombosis	213 (5.0)
Drugs to keep blood pressure up (Pressor)	3592 (84.3)
Cardiac balloon	467 (11.0)
Shock with STEMI	776 (18.2)
Unconscious with STEMI at cath lab	846 (19.9)
Shock at cath lab	1662 (39.0)
Pneumonia	1556 (36.5)
Intracranial hemorrhage	48 (1.1)
GCS Motor 1	3283 (77.0)
GCS Motor 2	114 (2.7)
GCS Motor 3	255 (6.0)
GCS Motor 4	223 (5.2)
GCS Motor 5	99 (2.3)
GCS Motor 6	60 (1.4)

B.2 LCA

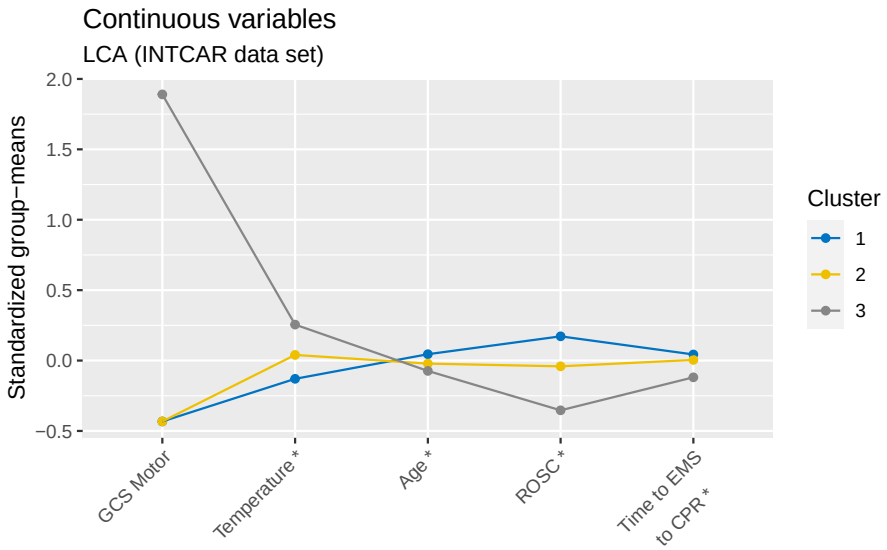

Figure 30: Distribution of the continuous variables of the INTCAR data set across different clusters as obtained by LCA. If there is a significant difference, the variable is marked by an asterisk, *.

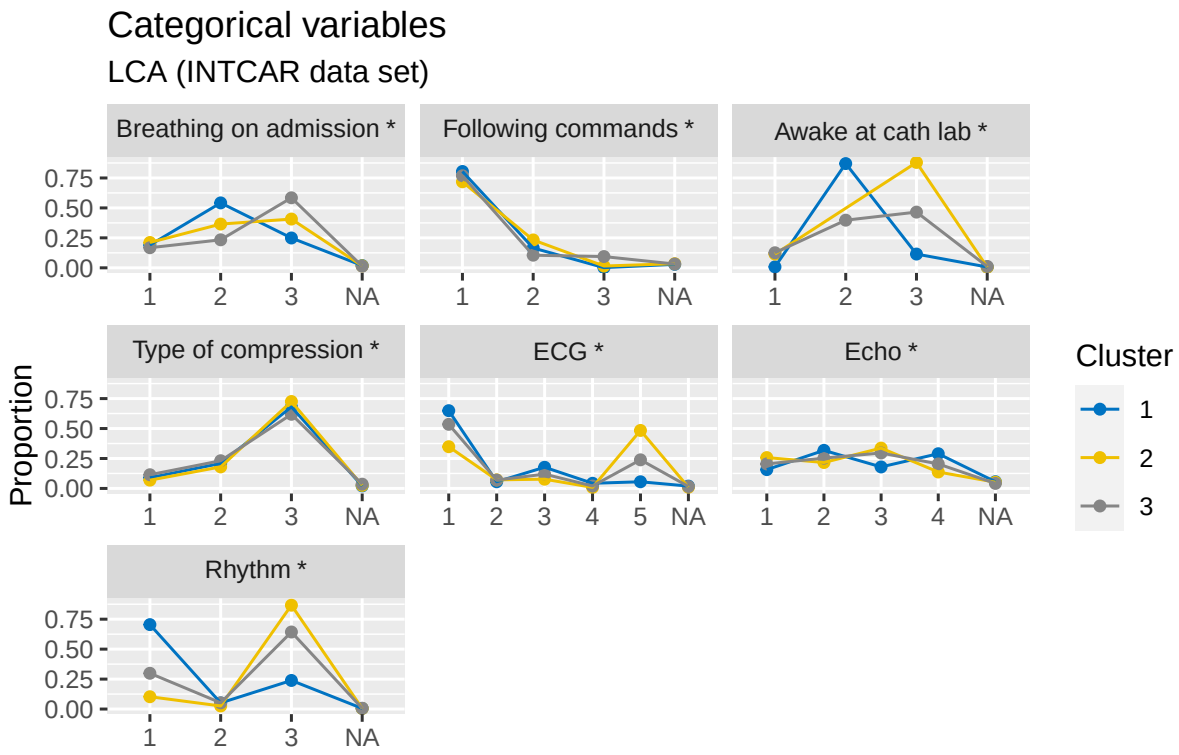

Figure 31: Distribution of the categorical variables of the INTCAR data set across different clusters as obtained by LCA. If there is a significant difference, the variable is marked by an asterisk, *.

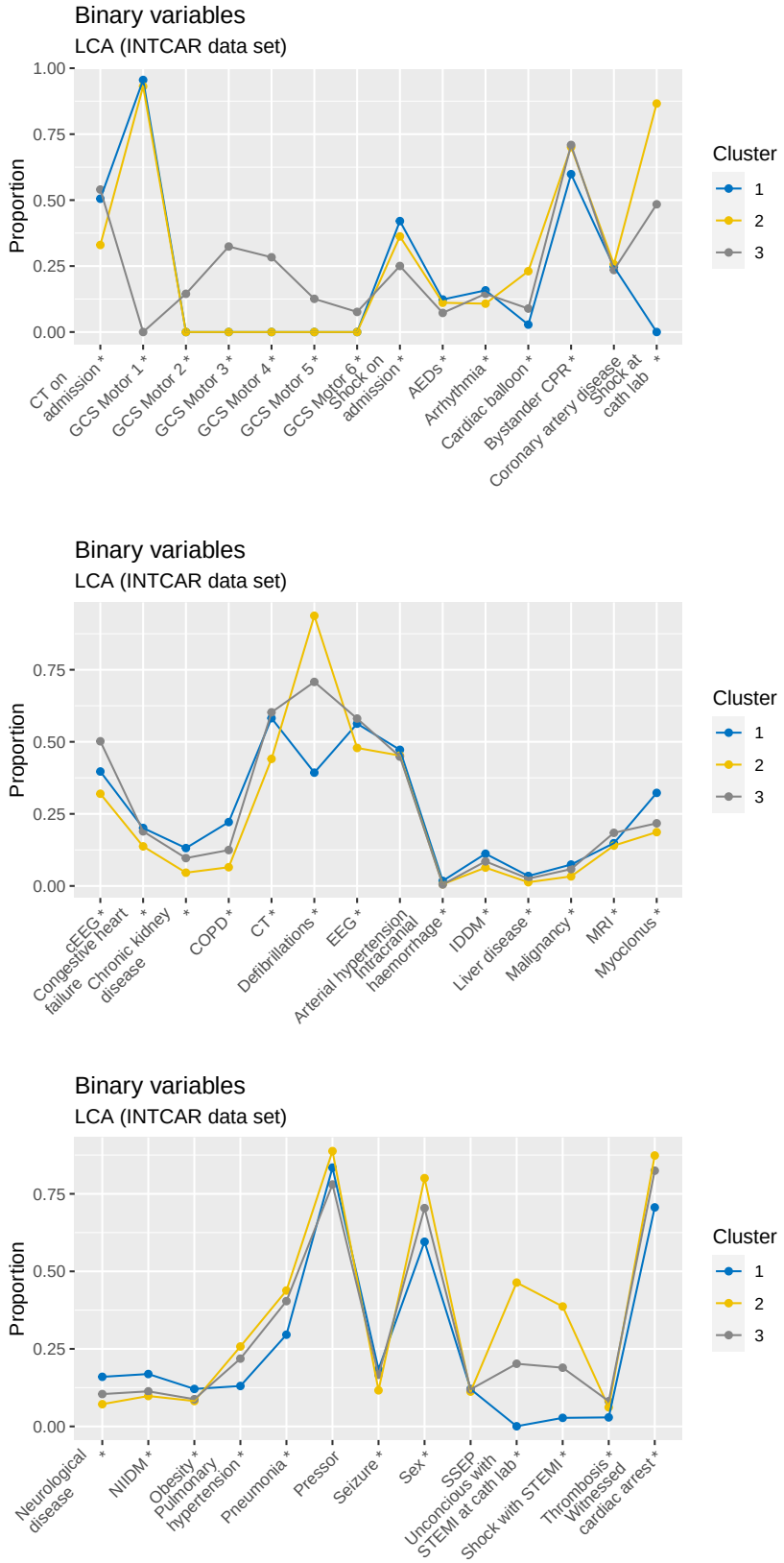

Figure 32: Distribution of the binary variables of the INTCAR data set across different clusters as obtained by LCA. If there is a significant difference, the variable is marked by an asterisk, *.

B.3 KAMILA

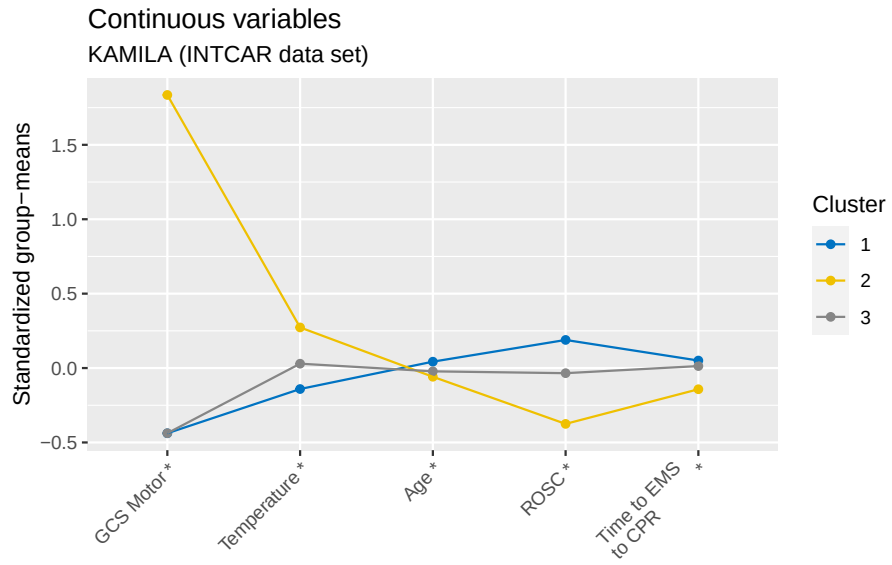

Figure 33: Distribution of the continuous variables of the INTCAR data set across different clusters as obtained by KAMILA. If there is a significant difference, the variable is marked by an asterisk, *.

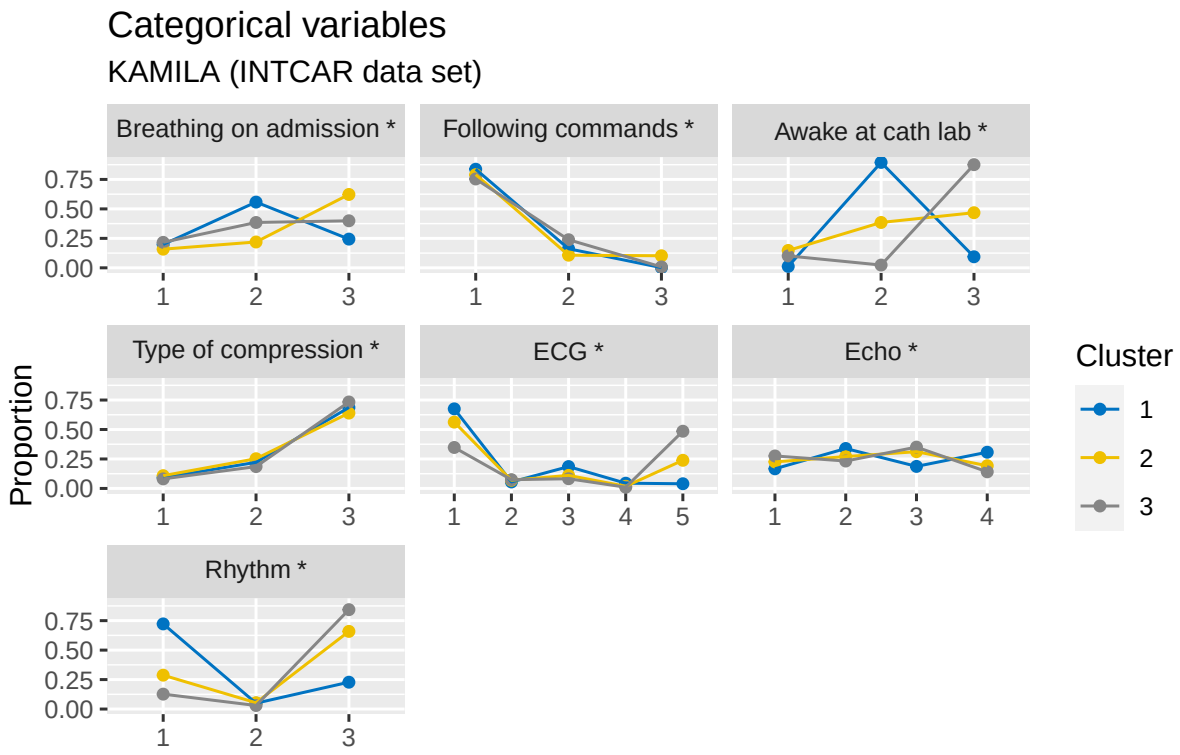

Figure 34: Distribution of the categorical variables of the INTCAR data set across different clusters as obtained by KAMILA. If there is a significant difference, the variable is marked by an asterisk, *.

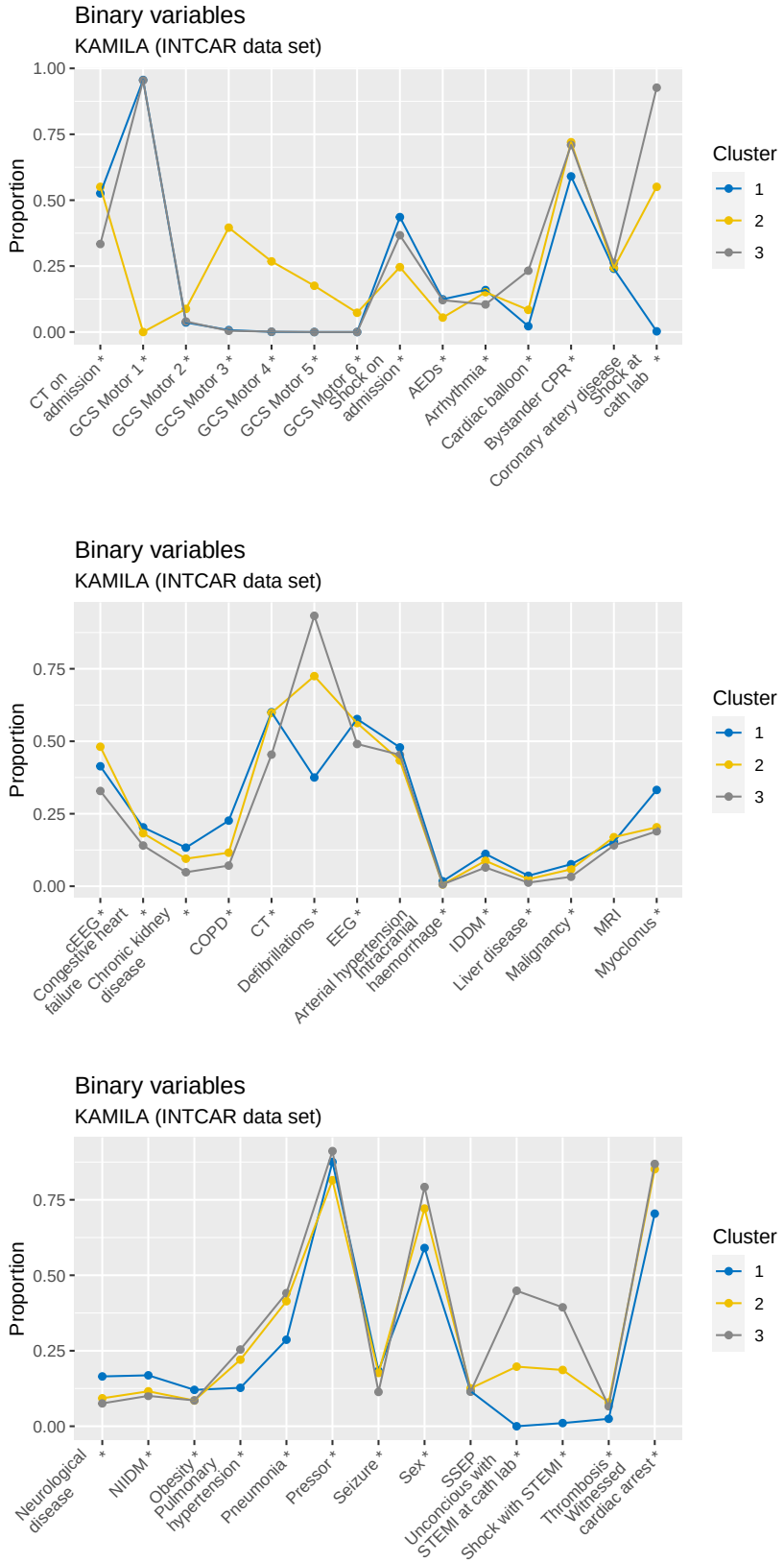

Figure 35: Distribution of the binary variables of the INTCAR data set across different clusters as obtained by KAMILA. If there is a significant difference, the variable is marked by an asterisk, *.

B.4 k-prototypes

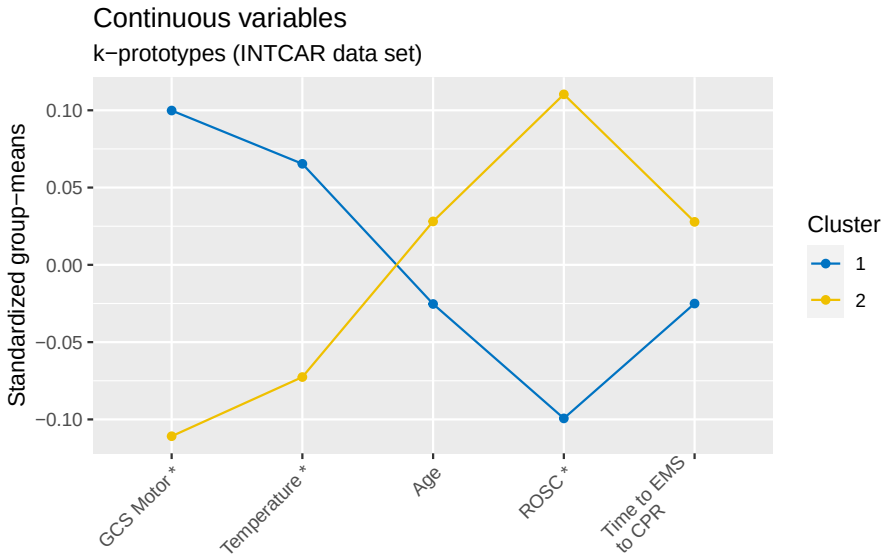

Figure 36: Distribution of the continuous variables of the INTCAR data set across different clusters as obtained by k-prototypes. If there is a significant difference, the variable is marked by an asterisk, *.

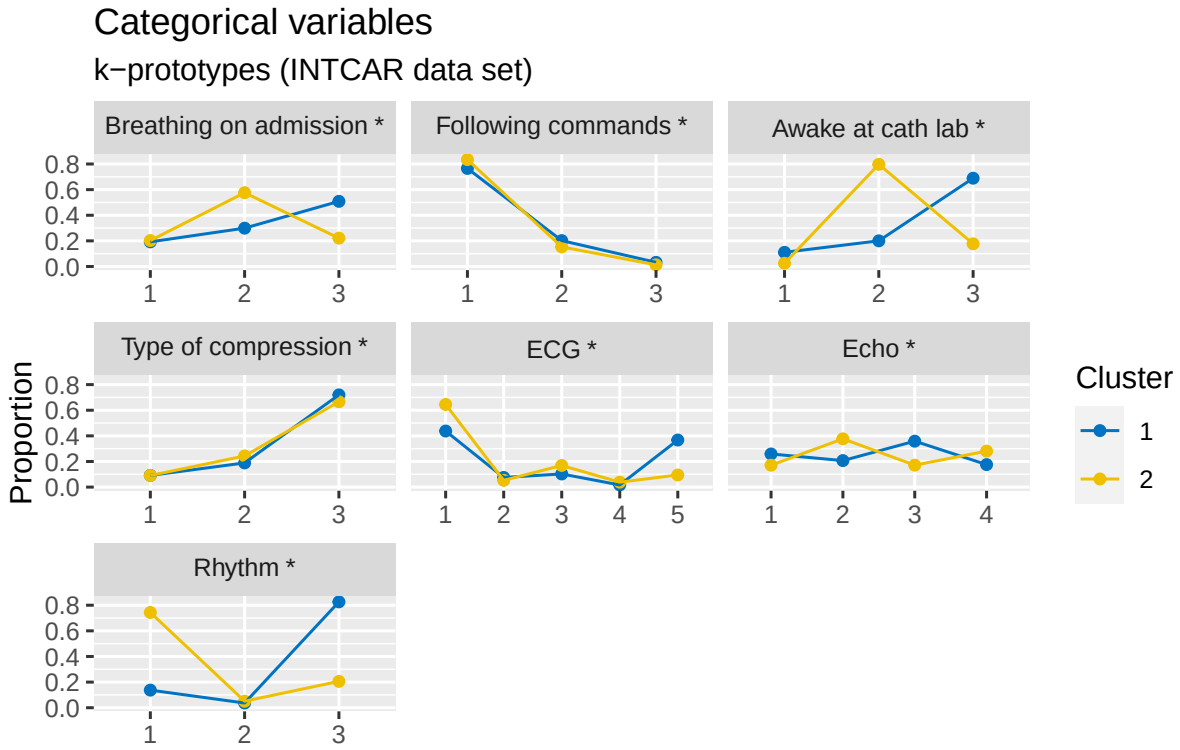

Figure 37: Distribution of the categorical variables of the INTCAR data set across different clusters as obtained by k-prototypes. If there is a significant difference, the variable is marked by an asterisk, *.

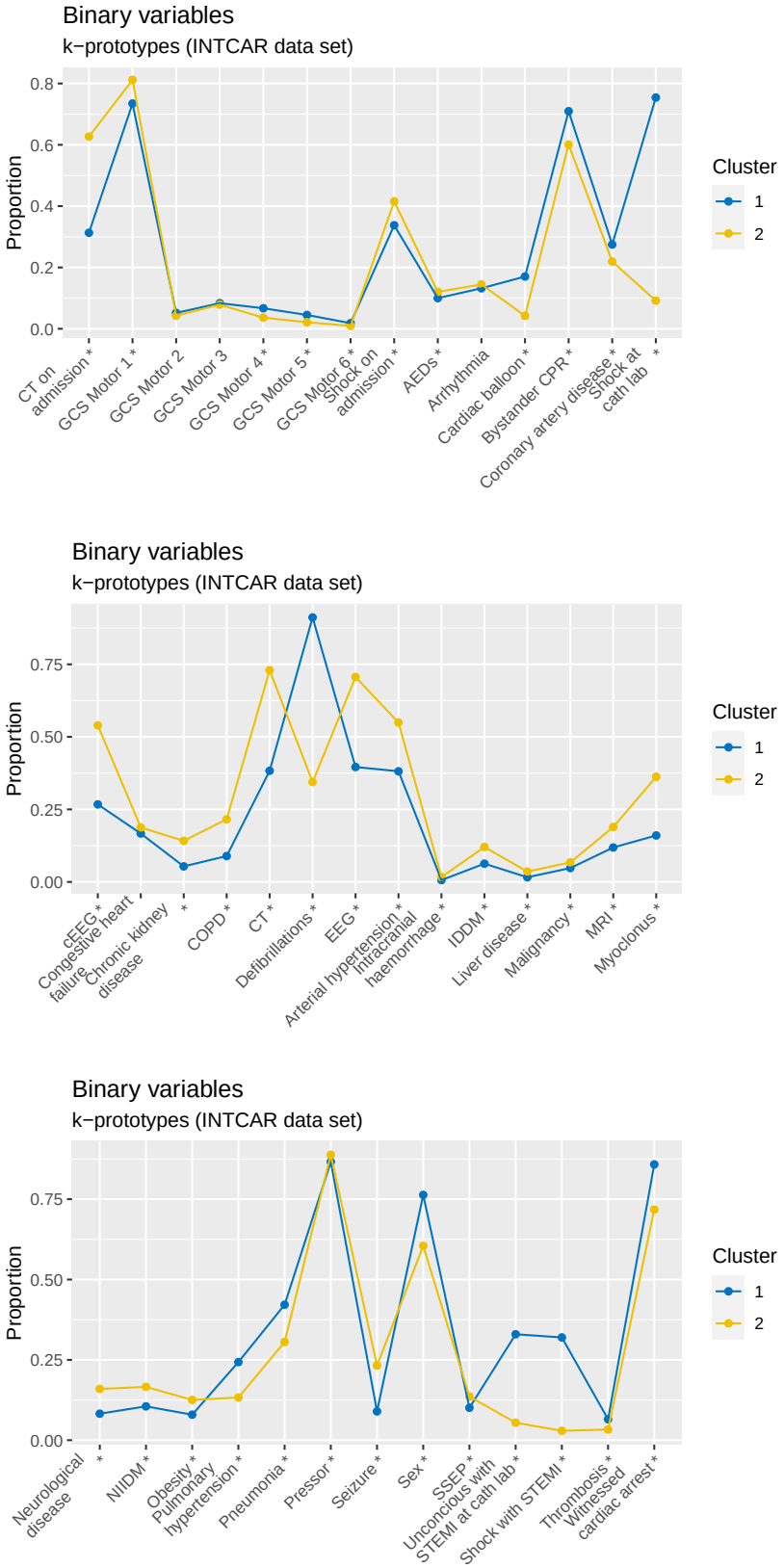

Figure 38: Distribution of the binary variables of the INTCAR data set across different clusters as obtained by k-prototypes. If there is a significant difference, the variable is marked by an asterisk, *.

B.5 PAM

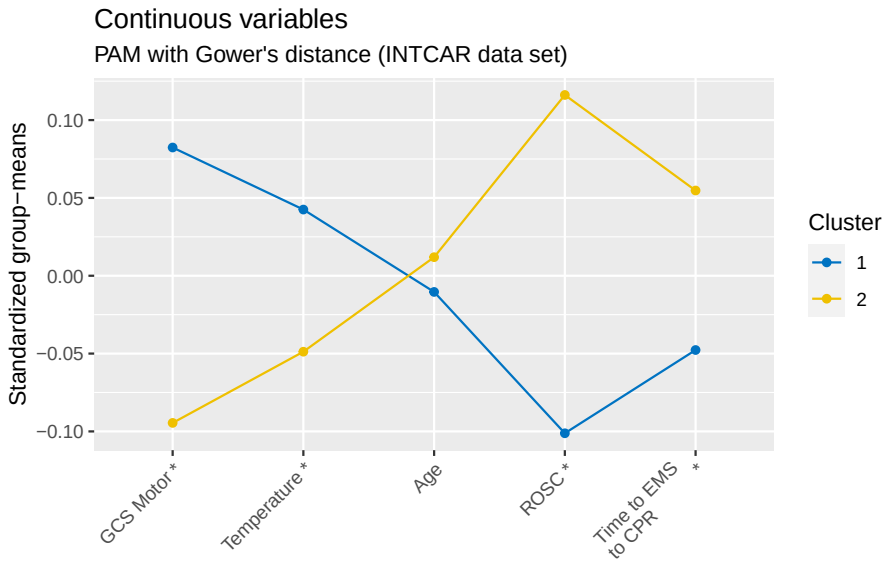

Figure 39: Distribution of the continuous variables of the INTCAR data set across different clusters as obtained by PAM with Gower's distance. A significant difference is marked by an asterisk, *.

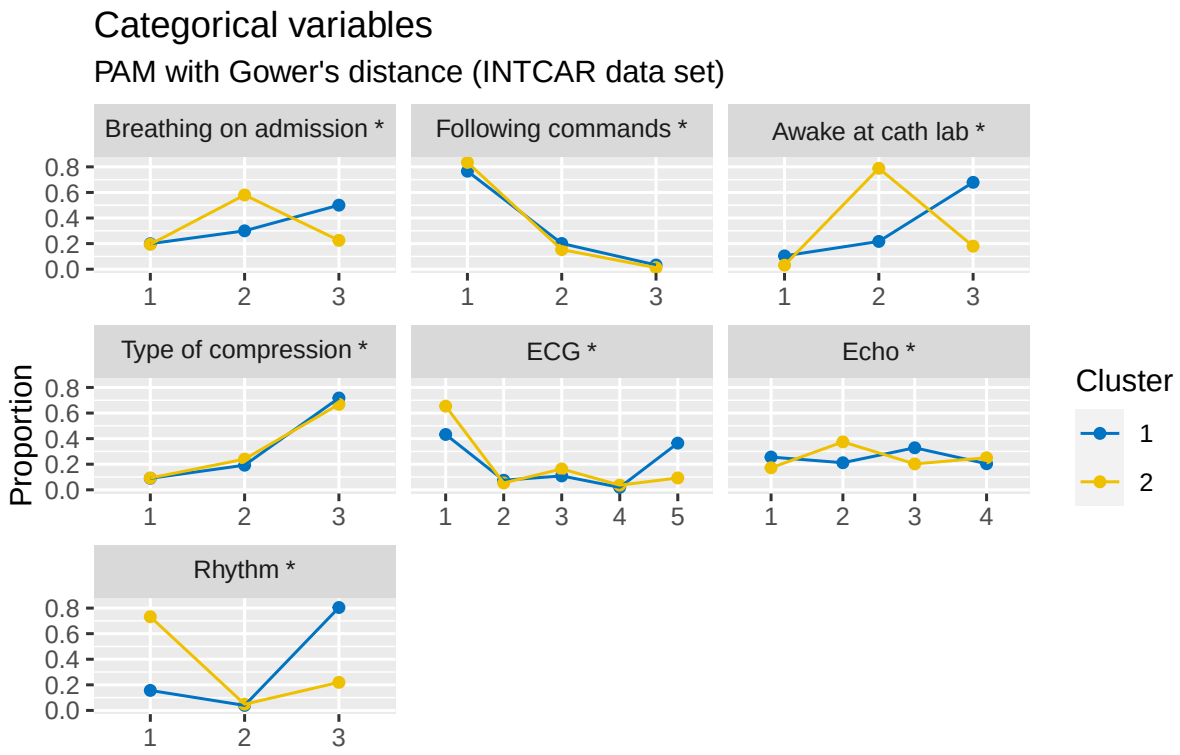

Figure 40: Distribution of the categorical variables of the INTCAR data set across different clusters as obtained by PAM with Gower's distance. A significant difference is marked by an asterisk, *.

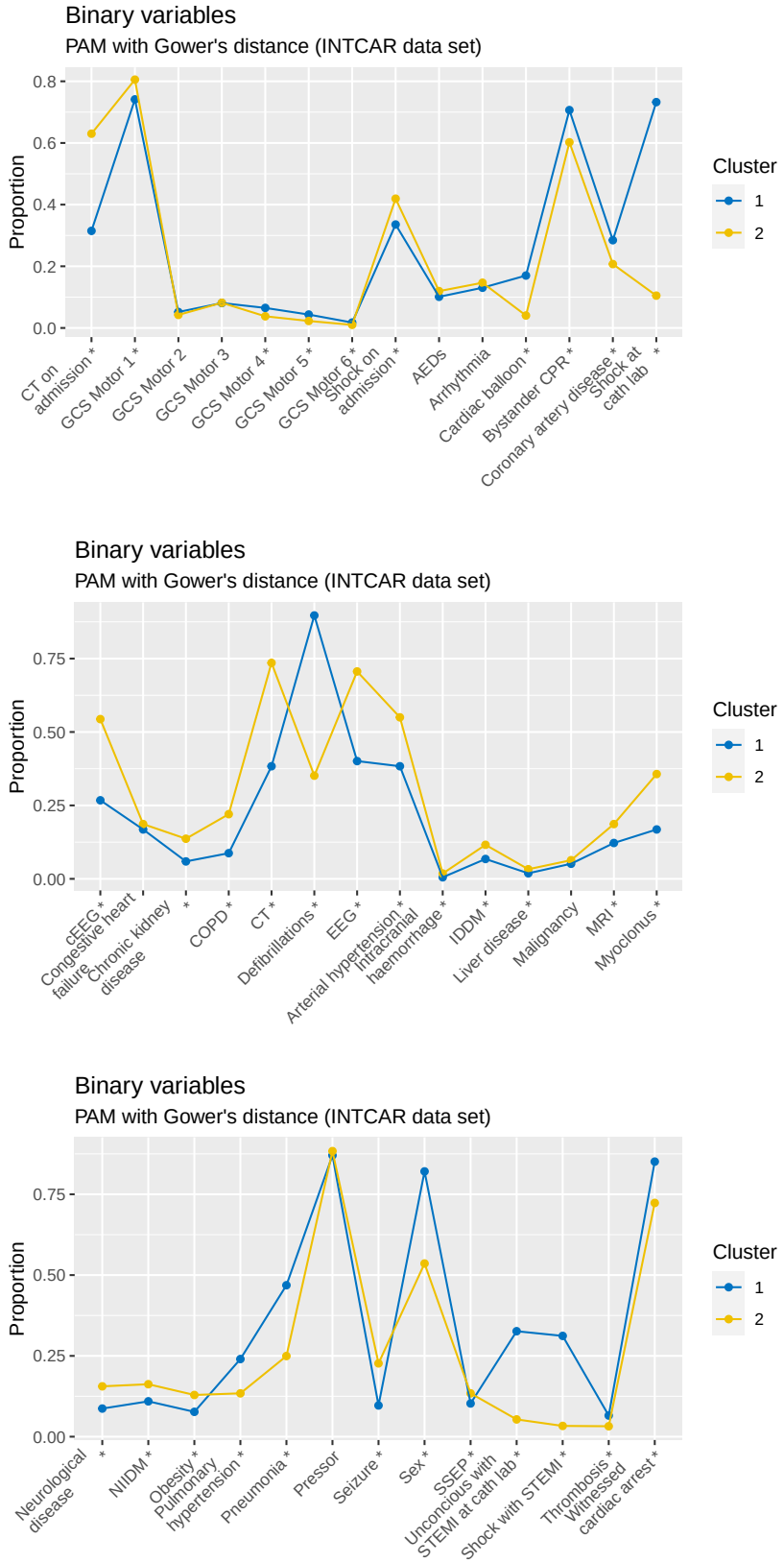

Figure 41: Distribution of the binary variables of the INTCAR data set across different clusters as obtained by PAM with Gower's distance. If there is a significant difference, the variable is marked by an asterisk, *.

Master's Theses in Mathematical Sciences 2021:E44
ISSN 1404-6342

LUNFMS-3099-2021

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/